Specification and Simulated Likelihood Estimation of a Non-normal Treatment-Outcome Model with Selection: Application to Health Care Utilization*

Partha Deb
Department of Economics
Hunter College, City University of New York

Pravin K. Trivedi†
Department of Economics
Indiana University

August 2004

Abstract

We develop a specification and estimation framework for a class of nonlinear, nonnormal microeconometric models of treatment and outcome with selection. A latent factor structure is used to accomodate selection into treatment and simulated likelihood methods for estimation. The methodology is applied to examine the causal effect of managed care on the utilization of health care services. The model consists of a discrete choice module for health insurance plans and an outcome equation with either binary or count dependent variable measuring utilization. The results indicate that there are significant unobserved self-selection effects and these effects substantially change the estimated effects of insurance on utilization.

Keywords: Endogenous treatment; latent factors; managed care; health care utilization.

JEL codes: C35; C15; I11.

*We gratefully acknowledge research support from the Agency for Health Research and Quality (R01 HS10904-02). We wish to thank David Zimmer for excellent research assistance. We also thank participants at the 11th European Workshop on Econometrics and Health Economics in Lund, Sweden, and especially Edward Norton, for their comments. We also thank Co-Editor John Rust and three anonymous referees for their comments and suggestions for improvement.

†Corresponding author: Department of Economics, Indiana University, Wylie Hall, Bloomington, Indiana 47405, USA. Email: trivedi@indiana.edu, Phone: +1 812-855-3567, Fax: +1 812-855-3736.

1. Introduction

In this paper we develop a specification and estimation framework for a class of nonlinear, nonnormal microeconometric models of treatment and outcome with selection. The primary focus in such models is usually on the effect of an endogenous treatment variable on an economic outcome. The model specification comprises of an outcome equation with a structural-causal interpretation and other equations that model the generating process of treatment variables. We apply our methodology to analyze the important but empirically unresolved question of the causal effect of managed care (treatment) on the utilization of health care services (outcome). Our work is related to that of Goldman (1995), Cardon and Hendel (2001) and Mello, et al. (2002) in that we also construct and estimate an econometric model that takes the endogeneity of insurance status into account. We distinguish between HMOs and other types of managed care plans which makes insurance a multinomial choice, while others have typically examined binary insurance choices.[1] We examine a comprehensive set of available measures of curative and preventive health care utilization using nationally representative data for the U.S.[2]

Our approach has numerous generic connections with the empirical microeconometric literature which are illustrated by the following selected examples. In labor economics, Bingley and Walker (2001) examine the effect of duration of husbands' unemployment on wives' discrete labor supply choices, Pitt and Rosenzweig (1990) study the effect of endogenous health status of infant children on their mothers' main daily activity and Carrasco (2001) examines the effect of childbirth on labor force participation of women. In treatment-outcome models related to fertility, Jensen (1990) examines the effect of contraceptive use on duration between births, while Olsen and Farkas (1989) examine the effect of childbirth on the hazard of dropping out of school. In health economics, Kenkel and Terza (2001) examine the effect of physician advice on the consumption of alcohol and Gowrisankaran and Town (1999) study the effect of hospital

[1] Cardon and Hendel (2001) also distinguish between HMO's and other forms of managed care.

[2] Goldman and Mello, et al. examine visits to the doctor and hospital stays; Cardon and Hendel examine total expenditures only. Goldman examines data from CHAMPUS (insurance for individuals who have served in the military), Cardon and Hendel use nationally representative data from 1987 and Mello, et al. examine the Medicare population from 1993-1996.

choice on the hazard of death in a hospital. In addition, Terza (1998) and van Ophem (2000) model the effect of household vehicle ownership on counts of trips.

These models, and ours, share several statistical features. First, both treatment and outcome processes are non-normal and nonlinear: multinomial, count, discrete or censored. Second, in each model the treatment is endogenous. Finally, investigators often have good a priori reasons for choosing particular (and uncontroversial) marginal models for treatments and conditional (on treatments) models for outcomes. But, the transition from given marginal and conditional distributions to a joint model for treatment and outcome continues to be a methodological challenge because typically nonnormal multivariate distributions are involved. In some cases, the marginal models have no (or very restrictive) tractable multivariate counterparts (e.g., in models of counts and durations). In others, treatment and outcome are from different statistical families (e.g., treatment being a multinomial and the outcome being a hazard rate) and so analytically tractable multivariate distributions often do not exist.

We develop a treatment-outcome model for multinomial choice of health insurance plans and either counts or binary choices of health services utilization with a view to evaluating the treatment effect of health insurance choice on health care utilization outcomes. We specify a joint distribution of endogenous treatment and outcome using a latent factor structure. Latent factors are incorporated into the treatment and outcome equations to allow for idiosyncratic influences on insurance plan choice to affect utilization, thus enabling us to make a distinction between selection on unobservables and selection on observables (Heckman and Vytlacil, 2001). These idiosyncratic influences are interpreted as unobserved heterogeneity. The model captures heterogeneity in the utilization response to insurance plans, which is known to be an important feature of program impact evaluation studies. (Heckman, Smith and Clements, 1997).

The latent factor structure has three main advantages over alternative ways of generating correlated errors. First, they can be used to combine conditional and marginal distributions, appropriately chosen, into an appropriate joint distribution. Second, they have a natural interpretation as proxies for unobserved covariates since they enter into the equations in the same way as observed covariates. The factor loadings can therefore be interpreted in much the same way as coefficients on observed covariates can.

Third, they provide a parsimonious representation of error correlations in models with large numbers of equations. In a later section we compare our approach with some alternatives.

We apply maximum simulated likelihood (MSL) techniques to estimate the parameters of our models. Simulation is used to evaluate integrals in the likelihood function of the model as no closed form solutions exist. Because of the complexity of our model, standard simulation methods are quite slow. Therefore, we adapt an acceleration technique that uses quasi-random draws based on Halton sequences (Bhat, 2001; Train, 2002).

The remainder of this paper is organized as follows. In section 2, we provide some background and an estimation framework for our empirical application. In section 3, we describe the simultaneous equations model and the estimation methods. We describe the data in section 4 and analyze the empirical results in section 5. Section 6 concludes.

2. Background

In the context of the market for health insurance in the United States, managed care plans seek to provide cost-effective care by using a variety of financial and nonfinancial tools to manage the use of health care services. Managed care plans in the market place include those offered by health maintenance organizations (HMO), preferred provider organizations (PPO), and point of service (POS) plans. HMOs typically are health plans with networks of providers, require enrollees to sign up with a primary care provider (PCP) and who do not pay for out-of-network services. PPO plans typically have a network of providers but do not require a signup with a primary care provider (PCP) and do pay for out-of-network services. Most POS plans are likely to have physician networks and PCP signup requirements and also pay for out-of-network services. But there is variation within these broad categories. For example, Dranove (2000) mentions four categories of HMOs based on how the physicians are organized. These include staff model, group model, independent practice association, and network model. Staff model HMOs are insurance companies that employ their physicians and pay them salaries, thereby eliminating the direct financial incentive to generate additional demand for their

services and fostering patient moral hazard. Group model HMOs are insurance companies that enter into exclusive contractual arrangements with large physician groups to provide professional services. Both staff and group model HMOs perform a gatekeeping role by controlling referrals to specialist services and hospitals that are more expensive than office visits to HMOs. Service control mechanisms include selecting a network of providers, deemphasizing specialist care while relying on primary care, using primary care physicians as gatekeepers to specialist and other care, using financial incentives to encourage cost containment, and so forth. Among these, plans with gatekeepers have the most direct provider-side control on the use of services by consumers. Gatekeeping is an identifying feature of HMO, PPO and POS plans. Although gatekeeping was hailed as the solution to the problem of moral hazard in the early years, it has been recently demonized in the popular press and in public opinion for being too restrictive. Nevertheless, they continue to be an important feature of managed care plans with tight utilization controls.

Studies have shown that HMO's are associated with lower hospitalization rates, reduced lengths of stay, the same or more office visits, and greater use of preventive services (for reviews, see Miller and Luft, 1994; Glied, 2000). But many of these studies do not control adequately for selection into HMO's, if at all. At least three forms of selection bias plague studies of incentives. First, self-selection arises because optimizing individuals, possessing knowledge of their own health attributes, proclivities, and economic constraints, select plans accordingly. Self-perceived healthy individuals, expecting on average lower demand for future health care, may choose low-cost plans with fewer choices than their less healthy and less constrained counterparts. Others may have preferences for certain modes of care, e.g., office-based care from their family physician, and hence may choose plans with generous benefits in that dimension. Therefore these attributes which partly determine the individual's choice of health plans also affects their expected utilization of services. This type of self-selection bias is emphasized in this article.

Second, self-selection into plans can arise from the economic behavior of health plans (Frank, et al., 2000; Cao and McGuire, 2002). Health plans that are offered by employers are often paid mostly through capitation or fixed payments. In such cases,

profit-oriented health plans have an incentive to distort the quality of services they offer to attract profitable and deter unprofitable enrollees. For example, if demand for treatment of expensive chronic conditions is better anticipated and more unevenly distributed in a population than demand for less expensive acute care, then the health plan has an incentive to distort the mix of its care away from chronic care and towards acute illness in order to deter the high risks and attract the low risks. Frank, et al. (2000) show how the incentives to distort services depend in a relatively straightforward way on means and correlations among predicted values of health care services in a population. In an empirical analysis, they find that if people are assumed to know a few of their own relevant characteristics (age, sex and prior spending) selection incentives can be quite severe.

Finally, self-selection can arise via the actions of health care providers. Dranove (2000) suggests that those providers with less aggressive treatment styles may be relatively more probable participants in capitated plans and salaried employment. In relation to the utilization of health care services the two sources of self-selection may be mutually reinforcing, there being some evidence that enrollees of HMOs with capitated or salaried providers receive less medical care than do the patients of fee for service providers (Hillman, Pauly, and Kerstein, 1989).

An econometric implication of the foregoing discussion is that there is strong a priori justification for treating insurance choice as endogenous and jointly determined along with health care utilization. In most survey data, the issue of self-selection bias cannot be resolved simply by including control covariates. The relatively limited evidence on selection on unobservables suggests that such selection effects on utilization of health care services can be substantial (Goldman, 1995; Mello, et al., 2002).

3. Econometric Methods

We begin by presenting a general representation of our model which has two modules, a choice-of-plan module and a utilization module. The modules are linked because plan choices are regressors in the utilization module and because there are common unobservable (latent) factors. Then we discuss the specific parametric forms and distri-

butional assumptions. Finally, we present and discuss pros and cons of some alternative approaches.

3.1. Model Specification

Let y_i^* denote the value of the latent (optimum) variable underlying the observed values of utilization, y_i. Let EV_j^* denote the (latent) indirect utility associated with the j^{th} insurance plan, with $j = 0, 1, 2$ corresponding to plans without gatekeepers (NMC), managed care plans that involve gatekeepers but are not HMO's (OMC), and HMO's (HMO), respectively. Let d_j be binary variables representing the observed choices. We treat NMC as the baseline choice.

The outcome or utilization equation for individual i, $i = 1, ..., N$, is formulated as

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \sum\nolimits_j \lambda_j l_{ji} + \varepsilon_i \tag{1}$$

where $\mathbf{x}_i$ is a set of exogenous covariates and $\boldsymbol{\beta}$, γ_1, and γ_2 are parameters associated with the exogenous covariates and insurance dummy variables. The error term is partitioned into ε_i, an independently distributed random error, and latent factors l_{ji} which denote unobserved characteristics common to individual i's choice of insurance plan of type j and health services utilization of that individual. The λ_j, factor loadings, are parameters associated with the latent factors.

The transformation from y_i^* given in (1) to the observed random variable y_i is through an appropriate distribution function $\mathbf{f}$ such that

$$\Pr(Y_i = y_i | \mathbf{x}_i, d_{1i,} d_{2i}, l_{ji}) = \mathbf{f}(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \sum\nolimits_j \lambda_j l_{ji}). \tag{2}$$

Following the random utility framework (McFadden, 1980, p. S15), the indirect utility or propensity to select insurance plan j, $j = 0, 1, 2$, is formulated as

$$\text{EV}_{ji}^* = \mathbf{z}_i'\boldsymbol{\alpha}_j + \delta_j l_{ji} + \eta_{ji}. \tag{3}$$

where $\mathbf{z}_i$ denotes exogenous covariates, α_j the associated parameters and η_{ji} are random error terms assumed to be independent of ε_i. Once again, l_{ji} are latent factors and δ_j are associated factor loadings. The transformation from the latent variable formulation

to the observed choices is via a distribution function $\mathbf{g}$ that describes a multinomial choice model such that

$$\Pr(d_{ji} = 1|\mathbf{z}_i, l_{ji}) = \mathbf{g}(\mathbf{z}_i'\boldsymbol{\alpha}_j + \delta_j l_{ji}), \quad j = 0, 1, 2. \tag{4}$$

We denote covariates in this plan-choice module by $\mathbf{z}$ and covariates in the utilization module by $\mathbf{x}$ to highlight the fact that they contain different variables in the empirical analysis. These issues are presented after the data are described.

Because the latent factors l_{ji} enter both insurance choice (4) and utilization (2) equations, they capture the individual-specific (or idiosyncratic) factors that induce self-selection into insurance plans through unobservables on utilization of health care services. Observe also that such a specification can explicitly incorporate heterogeneity in the response of utilization to insurance plan. From a statistical perspective, $\left(\delta_j l_{ji} + \eta_{ji}\right)$ and $\left(\sum_j \lambda_j l_{ji} + \varepsilon_i\right)$ are correlated composite error terms, even though $\left(\eta_{ji}, \varepsilon_i\right)$ are uncorrelated.

Under these assumptions, the joint distribution of selection and outcome variables, conditional on the common latent factors, can be written as

$$\begin{aligned} \Pr(Y_i = y_i, d_{ji} = 1|\mathbf{x}_i, \mathbf{z}_i, l_{ji}) &= \mathbf{f}(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \textstyle\sum_j \lambda_j l_{ji}) \\ &\quad \times \mathbf{g}(\mathbf{z}_i'\boldsymbol{\alpha}_j + \delta_j l_{ji}) \end{aligned}. \tag{5}$$

The problem in estimation arises because the l_{ji} are unknown. Although the l_{ji} are unknown, we assume that the distribution of l_{ji}, $\mathbf{h}_j$, is known and can therefore be integrated out of the joint density, i.e.,

$$\begin{aligned} \Pr(Y_i = y_i, d_{ji} = 1|\mathbf{x}_i, \mathbf{z}_i) &= \int \Big[\mathbf{f}(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \textstyle\sum_j \lambda_j l_{ji}) \\ &\quad \times \mathbf{g}(\mathbf{z}_i'\boldsymbol{\alpha}_j + \delta_j l_{ji})\Big] \mathbf{h}_j(l_{ji}) dl_{ji} \end{aligned}. \tag{6}$$

Cast in this form, the unknown parameters of the model may be estimated by maximum likelihood.

The maximum likelihood estimator solves the following problem:

$$\arg \max_{\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}} \prod_{i=1}^{N} L\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | y_i, d_{ji}, \mathbf{x}_i, \mathbf{z}_i\right), \tag{7}$$

where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \gamma_1, \gamma_2, \boldsymbol{\lambda})$ and $(\boldsymbol{\theta}_2 = \boldsymbol{\alpha}_j, \boldsymbol{\delta}_j, j = 0, 1, 2)$ refer to parameters in the outcome and plan choice equations respectively, and L refers to the joint likelihood whose i^{th} component is defined in (6).

The main computational problem, given suitable specifications for $\mathbf{f}$, $\mathbf{g}$ and $\mathbf{h}_j$, is that the integral (6)does not have, in general, a closed form solution. But this difficulty can be addressed using simulation-based estimation (Gourieroux and Monfort, 1996) by noting that

$$\begin{aligned} \Pr(Y_i = y_i, d_{ji} = 1 | \mathbf{x}_i, \mathbf{z}_i) &= \mathrm{E}\Big[\mathbf{f}(x_i'\boldsymbol{\beta} + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \textstyle\sum_j \lambda_j l_{ji}) \\ &\quad \times \mathbf{g}(z_i'\boldsymbol{\alpha}_j + \delta_j l_{ji})\Big] \\ &\approx \frac{1}{S}\sum_{s=1}^{S}\Big[\mathbf{f}(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \textstyle\sum_j \lambda_j \widetilde{l}_{jis}) \\ &\quad \times \mathbf{g}(z_i'\boldsymbol{\alpha}_j + \delta_j \widetilde{l}_{jis})\Big] \end{aligned}, \tag{8}$$

where $\widetilde{l}_{jis}$ is the s^{th} draw (from a total of S draws) of a pseudo-random number from the density $\mathbf{h}_j$ and $\widetilde{\Pr}$ denotes the simulated probability. A simulated likelihood function for the data can then be defined. The MSL estimator maximizes the average simulated log likelihood. Provided that S is sufficiently large, the precise number being a function of N, the maximization of the simulated likelihood is equivalent to maximizing the likelihood. The use of a fixed value of S as N increases results in the MSL estimator being biased. The literature indicates that S should increase faster than $\sqrt{N}$, but this does not give explicit guidance in choosing S. We discuss issues of simulation in greater detail below. Because of the complexity of our model, standard simulation methods are quite slow. Therefore, we adapt an acceleration technique that uses quasi-random draws based on Halton sequences (Bhat, 2001; Train, 2002). These methods are described in detail in Appendix A.

We maximize the simulated likelihood using a quasi-Newton algorithm requiring only first derivatives. Post-convergence the variance of the MSL estimates is obtained using the robust 'sandwich' formula for two reasons. First, information matrix and outer product formulae are inappropriate because they do not take into account uncertainty due to simulation chatter (McFadden and Train, 2000). Second, the appropriate variance formula under the quasi-likelihood interpretation is the robust form (White, 1982).

3.2. Parametric Forms and Normalization

The outcome variable in our model is either a non-negative count or a binary variable. For counts $y_i = 0, 1, 2, \ldots$, we specify $\mathbf{f}$ as the negative binomial-2 density,

$$f(y_i|\mu_i) = \frac{\Gamma(y_i + \psi)}{\Gamma(\psi)\Gamma(y_i + 1)} \left(\frac{\psi}{\mu_i + \psi}\right)^{\psi} \left(\frac{\mu_i}{\mu_i + \psi}\right)^{y_i}, \tag{9}$$

where the conditional mean parameter $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \sum_j \lambda_j l_{ji})$ denotes the mean component of utilization and $\psi \equiv 1/\nu$,$(\nu > 0)$ is an overdispersion parameter in the conditional variance $\mu_i\,(1 + \psi\mu_i)$.

For the case where y_i is binary we specify $\mathbf{f}$ as the normal distribution (leading to the Probit model).

The functional form of the choice equations depends upon the stochastic assumptions about the random error in the utility function. In this paper, we assume that $\mathbf{g}$ has (conditionally on the latent factors) a mixed multinomial logit structure (MMNL) defined as

$$\Pr[d_{ji} = 1|\mathbf{z}_i, l_{ji}] = \frac{\exp(\mathbf{z}_i'\boldsymbol{\alpha}_j + \delta_j l_{ji}).}{\sum_{k=0}^{J} \exp(\mathbf{z}_i'\boldsymbol{\alpha}_k + \delta_k l_{ki})} \tag{10}$$

with the normalization restriction $\alpha_0 = 0$ and $j = 0, 1, 2$. The main goal of MMNL structure is to provide "good" estimates of plan choice probabilities.

The $\mathbf{h}_j$ are assumed to be standard normal densities. The zero mean assumption is without loss of generality and fixed variance is needed because the variance of the latent factors cannot be separately identified.

Since $\delta_0 = 0$ and $\alpha_0 = 0$ are required normalizations in the multinomial logit model, we assume $l_{0i} = 0$ without loss of generality, i.e., l_{1i} and l_{2i} are interpreted as factors favoring OMC and HMO to NMC. In addition, because the variances of the unobserved factors cannot be identified, a normalization is required on either λ_j or δ_j. We assume $\delta_j = 1$ for each j and treat λ_j as free parameters. [3]

[3] In preliminary work, we estimated models using the normalization $\lambda_j = 1$ while estimating δ_j. The substantive results in either case were identical.

3.3. Comparison with Alternative Estimators

The approach of this paper is based on an initial specification of the marginal - conditional models for outcome and treatments that are connected via unobserved heterogeneity. Another approach that begins with marginals and combines them into a joint model is the copula approach (Joe, 1997). The copula approach has been used specifically to estimate the selectivity models (Lee, 1983; van Ophem, 2000; Prieger, 2002) for nonnormal data. Given the choice of a copula from amongst a number of forms, the approach has the attraction of leading to a closed form model that can be estimated without resorting to simulation. However, depending upon the exact choice of a particular functional form out of a number that have been described in the literature, the copula places restrictions on the pattern of allowable correlations. Some only permit positive correlations, and most place bounds on the permissible values. By contrast, our specification permits both negative and positive correlations, although it too places bounds on the correlation. In addition, extension of copulas to marginals for discrete random variables in different families of distributions is still in its infancy.

We have specified the error structure of our model using a factor-loadings specification. One could, however, use a general multivariate normal specification of unobserved heterogeneity. One advantage of the multivariate normal specification is that it imposes no bounds on the correlations. Another is that many might find the setup more familiar. However, there are two major disadvantages associated with specifying a multivariate normal distribution. First, one is forced to specify individual errors as being normally distributed, or one must choose from a very small set of alternative distributions for which the joint distributions are of known form. Second, in such 'multivariate' formulations, analytical derivatives with respect to correlation parameters are very difficult to specify and this imposes a large computational marginal cost.

It is also possible to specify the error structure of our model using discrete distributions as described by Mroz (1999). There are two advantages to such an approach. First, because such models are finite mixture models, they are semiparametric and the discrete distributions can, in principle, approximate any continuous distributions. Second, because such models replace the integration with addition, their likelihood functions

are considerably simpler to compute. There is one drawback of such a specification, one which we encountered frequently when we attempted to estimate such models in preliminary analyses. Because the discrete factor model is a finite mixture model, its likelihood function is known to have multiple maxima (Lindsay, 1995). In our application, we encountered this problem on a sufficiently frequent basis to make us stop using this specification, especially as estimates from alternative runs often gave effects of different signs, significance and magnitude. Note that Mroz (1999) does not report such parameter instability. But his evaluations of the method are in the context of a special case of a bivariate system with a binary treatment and a linear outcome.

It is also possible to construct a limited information maximum simulated likelihood estimator with desirable statistical and computational properties. One such limited information approach involves holding the parameters of the treatment equations fixed while estimating only the parameters of the outcome equation. Formally this is defined as two-step sequential estimator that solves:

$$\arg\max_{\{\boldsymbol{\theta}_1\}} \prod_{i=1}^{N} SL_{2S}\left(\boldsymbol{\theta}_1 | y_i, \mathbf{d}_{ji}, \mathbf{x}_i, \mathbf{z}_i, l_{ji}, \widetilde{\boldsymbol{\theta}}_2\right). \tag{11}$$

where SL denotes simulated likelihood, l_{ji} denotes a draw from the standard normal density, and

$$\widetilde{\boldsymbol{\theta}}_2 = \arg\max_{\{\boldsymbol{\theta}_2\}} \prod_{i=1}^{N} SL\left(\boldsymbol{\theta}_2 | \mathbf{d}_{ji}, \mathbf{z}_i, l_{ji}\right), \quad (j = 0, 1, 2) \tag{12}$$

defines the maximum simulated likelihood estimates for the parameters in the mixed multinomial logit (MMNL) plan selection equations based also on draws from the standard normal distribution.

This sequential estimator has the following interpretation and justification. The simulated likelihood at the second step is a weighted likelihood in which the weights are estimated probabilities generated by the first stage MMNL estimates. When the joint likelihood is maximized, the estimated probability weights are estimated simultaneously, not sequentially. The MMNL model for plan choices is analogous to a partial reduced form and outcome equation is a structural-causal equation. The parameters $\boldsymbol{\theta}_2$ are functionally independent of $\boldsymbol{\theta}_1$ and are analogous to nuisance parameters. The main

role of the MMNL component is to generate "good" estimates of the choice probabilities, or probability weights. The second step of the sequential estimator defined in (11) is analogous to maximization of a concentrated simulated likelihood function. Provided that the first step yields consistent estimates of $\boldsymbol{\theta}_2$, conditioning on them in the second step will yield consistent estimates of $\boldsymbol{\theta}_1$.

Although the sequential 2-step estimators do not appear to have been used in the context of the MSL estimation, they are quite natural and potentially very useful especially in large samples where the computational burden is very significant even with fast computers. An analysis of the properties of two-step estimators in the maximum likelihood framework is given in Pagan (1986) who provides conditions under which the two-step estimator is consistent and efficient. The efficiency of the two-step estimator is shown to depend crucially on the block diagonality of the asymptotic information matrix of parameters $\mathbf{I}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Here there are no a priori arguments to justify or reject this restriction.[4] We implement such an estimator as a check of the robustness of our results.

4. Data

In this study, we use data from the 1996 Medical Expenditure Panel Survey (MEPS). MEPS is a representative survey of the noninstitutionalized population in the U.S. with wide scope and excellent information on demographic characteristics, health status, employment status and earnings, and a wide variety of measures of health care utilization. Unfortunately, although initially designed to be a long panel, the final design of MEPS is a essentially a cross-section with measurements taken over two calendar years in 5 rounds of surveys. The 1996 sample consists of 21, 571 persons. In our study we focus on the subsample of non-elderly adults (ages 18 to 64) who have some form of private health insurance. In addition, we eliminate individuals who are covered by Medicaid or other public insurance plans and Medicare enrollees (both elderly and disabled) because

[4] The approach of this paper is non-Bayesian. Bayesian simulation-based methods are an attractive alternative to MSL. It is possible that they might be computationally more efficient. Models that might be computationally infeasible in the MSL framework either because of the large sample size, large number of parameters, or both, are often feasible to estimate using Bayesian methods (Geweke, et al., 2003).

we wish to focus on the role of HMO's and other managed care plans among persons who make such choices in the private market for health insurance. Thus we are left with a sample size of 8129.

4.1. Dependent Variables

Managed care is captured via two dummy variables, enrollment in an HMO (47%) and enrollment in other managed care plans, denoted OMC (8.2%). The remainder are in plans that do not have gatekeeper restrictions to care, denoted NMC. Enrollment status is measured at the first round of the survey in 1996.

Our empirical analysis covers five curative and five preventive measures of health care utilization, measured for the 1996 calendar year and verified from providers of care. The first set of curative utilization variables are frequencies of visits to different types of providers: to an MD in an office setting (including primary care physicians and specialists), to a non-MD medical professional in an office setting (including visits to psychologists and social workers, nurses and nurse practitioners), to a hospital, to the emergency room and to a hospital outpatient clinic ($N = 8129$). The second set of preventive care services are binary variables: whether blood pressure ($N = 7952$) and cholesterol checks ($N = 7717$) were received in the last two years, whether a flu shot was taken in the last year ($N = 7948$), and, for females only, whether a pap smear ($N = 4082$) and a mammogram ($N = 2105$) was received in the last year. Sample sizes for preventive care visits vary for two major reasons; some measures are restricted to one gender and others to specific age groups. In addition, some small differences are due to non-responses. Summary statistics for utilization are presented in Table 1. Relative to individuals in NMC plans, those in OMC have significantly higher doctor visits while those in HMO plans have significantly higher outpatient utilization. Persons in OMC and HMO plans are more likely to have received blood pressure checks and women in these plans are more likely to have received pap smears. Finally, those in HMOs are more likely to have their cholesterol checked than individuals in NMC plans.

4.2. Independent Variables and Exclusion Restrictions

Our choice of explanatory variables for the utilization and insurance choice equations is similar to that in Dowd, et al. (1991), Ettner (1997) and Goldman, et al. (1995). Socioeconomic characteristics include age, for which we have explored polynomial and linear spline specifications, gender, ethnicity, marital status, education, family size, location of residence, and annual personal income measured in the third round of the survey. Health characteristics include self-perceived health status, which we decompose into four dummy variables from the 5 point scale representing very good, good, fair and poor health (excellent health is the excluded category), the existence of a functional limitation and the number of chronic conditions. Self-perceived health status might be endogenous with respect to utilization of medical care. Because each of these health status measures is determined in the first round of the survey, they are predetermined in our study.

The determinants of insurance choice include all the socioeconomic and health characteristics that determine health care utilization. In principle, the parameters of the semi-structural model we have described are identified through nonlinear functional forms even if all the variables in the insurance equations are included in the utilization equation. However, for more robust identification we use traditional exclusion restrictions by specifying instrumental variables in the insurance choice module that are excluded from the utilization module. We use certain employment characteristics as identifying instruments: whether the individual is employed, whether the individual is self employed or works in the government sector, whether the individual belongs to a union, number of employees in the firm and whether it is in multiple locations, whether the person is employed in a blue collar or service job. Employment characteristics are used as instruments, conditional on a sample of individuals who are all covered by some form of insurance, because characteristics of firms are known to be important determinants of the supply of health plans available to individuals in the U.S. and that this is the main mechanism by which employment characteristics affect choice of health insurance plans (Johnson and Crystal, 2000). In addition, conditional on the socioeconomic and health characteristics we include as exogenous covariates in the insurance and uti-

lization equations, employment characteristics add little or no explanatory power to the utilization equation. Because there is no formal test for the validity of exclusion restrictions in a nonlinear setting such as ours, our checks of instrument relevance and exogeneity are informal.

As with most instruments that are not of an experimental or quasi-experimental nature, however, the validity of our choice of instruments may be questioned for a variety of reasons. First, there is evidence that employment and access to health insurance may be jointly determined (Gruber, 2000). By restricting our sample to only individuals who are covered by health insurance, we eliminate the issue of access to plans and thus minimize the issue of instrument validity along this dimension.[5] Second, it is possible that employment status, and self-employment status among those employed, may be jointly determined with the desire to have access to a particular type of health plan, even conditional on having health insurance and on observed health and socioeconomic characteristics. Meer and Rosen (2003) demonstrate that self-employment status is a valid instrument in a model of health insurance and utilization. Nevertheless, to investigate the possibility that the instruments may be invalid, as part of our robustness analysis we estimate models for subsamples of those who are employed and those who are employees of firms to check whether our baseline results change substantially or not. Finally, one may argue that individuals who work in certain sectors (e.g. government, service sector) may use more or less medical care than others either because they have preferences for job stability due to either poor health, or because industry health risks are bigger or smaller than in other sectors.[6] In either case, job sectors would be invalid instruments only if the person's own observed health status were inadequate, which we not believe to be the case. Although we do not investigate these possibilities in any formal way, we note that our results will show that sector of employment does not significantly determine choice of health insurance plans, conditional on being enrolled in a plan, thus reducing the force of this argument.

Descriptions and summary statistics of demographic, employment and health status

[5] Consequently, our results should be treated as identifying the causal effects of plan type conditional on having insurance coverage.

[6] We thank anonymous referees for raising these possibilities.

control variables stratified by insurance plan choice are presented in Table 2. Individuals enrolled in NMC plans have significantly different demographic characteristics than those enrolled in HMO plans and, although to a lesser extent, those who are enrolled in OMC plans. Employment characteristics are different too. Most noticeable are differences in firm size, measured both by number of employees ($firmsize$) and whether the firm is in one or more locations ($multlocation$). There are no statistically significant differences in observed health status measures. Although others have found differences in observed health status across insurance plan types (see, e.g., Mello, et al., 2002), these studies are about other populations and/or include the uninsured. Our sample consists largely of individuals who receive health insurance as an employment benefit.

5. Results

In this section we discuss the results from ten jointly estimated models. After some preliminary remarks regarding choice of simulation draws, we discuss the insurance choice equations. Then we discuss utilization, grouped into curative and preventive categories.

In complex nonlinear models such as ours, and especially those with large numbers of parameters, starting values of the parameters for the maximization algorithm can be critical for computational reasons. In our case the starting values were obtained by initially estimating the plan choice equations and the outcome equation under the restriction of exogenous choice dummies. In the work reported here we have used $S = 2000$ based on Halton draws. $S = 2000$ was chosen after considerable experimentation to determine the stability of the gradients of the likelihood function for different sets of simulation draws. Note that this is a considerably larger number than has been used in many empirical studies that use the MSL method. The most careful evaluations of numbers of draws required in simulation-based models are from the literature on multinomial choice models (see Train, 2002). In such models, the errors are correlated with each other because they form a seemingly unrelated system of equations. In our model, dependent variables from some equations enter the right hand side of other equations, thus the joint distributions of the errors are considerably more complex.

Although we have no formal proof, our experience estimating both types of models suggests that models with endogeneity require considerably more simulation draws than models that simply involve seemingly unrelated errors.

5.1. Health Insurance Choice

The estimates of the MMNL insurance equations from each of the ten models are very similar because they are all estimates for the same choices of type of health plans with the same sets of covariates. So we present and discuss estimates from only one of these models, that from the joint model of insurance and visits to the doctor.

Marginal effects from this model are presented in Table 3.[7] We find that older and rural individuals are more likely to choose NMC plans and less likely to choose OMC and HMO plans. Women and minorities are less likely to enroll in NMC plans and more likely to choose HMO plans. There are substantial regional differences as well. Health status indicators, educational attainment and income are generally not significant. These are reasonable results given that estimates are for a sample of individuals with private health insurance, most of whom obtain insurance from their employers (or from the employers of someone in the household). The insignificance of the health status variables in the choice equations suggests that for this particular population we do not have evidence of favorable selection on the basis of observed health status into HMOs. However, it is still possible that there is favorable selection on the basis of unobserved health status.

The insurance choice equations contain eight employment related variables that are excluded from the utilization equation. HMO enrollment is significantly, and positively related to being employed at a large firm ($firmsize$) with multiple locations ($multlocation$), and negatively related to $selfemployed$. Individuals who are self-employed or work for small firms are more likely to be enrolled in NMC plans. Employment sector and occupation are not significant determinants of choice of plan types.

[7] Marginal effects for continuous variables are calculated using appropriate derivatives from the outcome equation. Marginal effects for dummy variables are constructed using discrete changes in the expected outcome. Standard errors of the marginal effects are calculated using a Monte Carlo technique with 500 draws from the multivariate normal distribution with mean and covariance matrix set at the estimated MSL values.

These instruments are tested for joint significance in the MMNL using the likelihood ratio (LR) statistic and are statistically significant in each case. For example, the LR test statistic is 125 for the sample used to estimate the model for doctor visits and is 127 for the sample used to estimate the probability of blood pressure checks. Both values are relative to the conventional 95 per cent critical values for $\chi^2(16)$ and the result confirms that the instruments have useful predictive power and hence are statistically suitable identifiers.

5.2. Curative Health Care Services

Table 4 provides the estimated coefficients on the insurance dummy variables and the factor loadings associated with the latent factors for curative health care services. The full set of parameter estimates for the outcome equations is reported in Table 1 of the Appendix. The estimated coefficients are of plausible sign and significance. The coefficient of the HMO dummy variable is positive and highly significant for three measures: *Doctor*, *Outpatient* and *ER*. Thus, after correcting for self-selection, HMOs encourage the use of curative health care in a number of potentially cost-effective dimensions. Unfortunately, they also tend to promote the use of emergency room services, perhaps because they are treated as the primary mode of "after hours" care. The factor loading coefficient λ_{HMO} is estimated to be negative and highly significant in three equations (*Doctor*, *Outpatient* and *ER*) but λ_{OMC} is typically not. The interpretation of the significantly negative factor loading coefficient is that the unobserved factors that increase the probability of being enrolled in an HMO also lead to lower utilization relative to that of the randomly assigned HMO enrollee. This means that there is significant favorable selection on unobservables into the HMO plans.

Table 5 presents average treatment effects of HMO, i.e., $\mathrm{E}[y\,|\mathbf{x},d_j=1]-\mathrm{E}[y\,|\mathbf{x},d_j=0]$, and associated standard errors evaluated for a variety of hypothetical individuals.[8] For

[8] There is an extensive literature on the identifiability of this parameter under alternative estimation procedures. For example, Imbens and Angrist (1994) and Heckman and Vytlacil (2001) discuss the identification conditions. It is clear from these discussions that the widely used linear instrumental variable estimator identifies the ATE parameter only under additional restrictions such as common treatment effects (i.e, absence of heterogeneous treatment effects) or monotonicity of the effects. By contrast, our latent factor formulation can identify the ATE even when treatment effects are heterogeneous.

comparison, we have calculated both the effects from our joint model which accounts for endogeneity of plan-type and from the single-equation models which do not. Given the imprecise nature of the estimates on OMC coefficients, we do not report treatment effects with respect to OMC. The hypothetical individuals we consider have the average characteristics of the entire sample, of black individuals, of non-black individuals and of males and females. We also calculate treatment effects at the average characteristics of the sample of individuals with no chronic conditions and those with one or more chronic conditions. Finally, we calculate treatment effects at the median characteristics of individuals in the sample and the average characteristics of those actually enrolled in HMOs.

When endogeneity of plan-type is not accounted for, doctor visits are the only curative care with a statistically significant treatment effect. However, once self-selection is accounted for, doctor visits, outpatient visits and emergency room visits all have statistically significant treatment effects. For the individual with average characteristics and controlling for self-selection, those in HMOs are predicted to have 2.6 more doctor visits, 0.5 more outpatient visits and 0.13 more emergency room visits. In each case, the treatment impacts controlling for self-selection are much larger than the corresponding treatment effects assuming exogeneity. For example, when HMO status is treated as exogenous, an individual with "average" characteristics who is enrolled in an HMO has 0.25 more visits to the doctor. The corresponding effect when the endogeneity of HMO status is taken into account is 2.6.

The magnitudes of treatment effects obtained for the "average individual" are very similar to those obtained for characteristics set at the sample averages of individuals who are actually enrolled in HMOs, i.e., the "average treated individual". But the treatment effects for individuals who have median characteristics are substantially smaller, although statistically significant, than individuals with average characteristics. This demonstrates that the effect of being in an HMO differs substantially across individuals in the sample. The treatment effects are uniformly smaller for the average black individual as compared to the average non-black, for the average male as compared to the average female (except in the case of emergency room visits for which the treatment effects are very close) and for the average individual with chronic conditions as compared

to the average individual with no chronic conditions. These results collectively suggest that different groups of individuals react differently to the incentives and restrictions on care implied by managed care models of health care provision.

5.3. Preventive Health Care Services

Table 6 provides the estimated coefficients on the insurance dummy variables and the factor loadings associated with the latent factors for preventive health care services. The full set of parameter estimates for the outcome equations is reported in Table 2 of the Appendix. The estimated coefficients are of plausible sign and significance. The coefficient of the HMO dummy variable is positive and highly significant for three measures: $Bloodpressure$, $Cholesterol$ and $Flushot$. In addition, it is positive and marginally significant for $Mammogram$. In general, after correcting for self-selection, HMOs encourage the use of preventive health care. For OMC enrollees the evidence is weak and statistically insignificant, except in the case of $Mammogram$ where it is negative and it is statistically significant. The factor loading coefficient λ_{HMO} is estimated to be negative and highly significant for $Bloodpressure$, $Cholesterol$ and $Flushot$, but λ_{OMC} is typically not. Once again, the interpretation of the significantly negative factor loading coefficient is that the unobserved factors that increase the probability of being enrolled in an HMO also lead to lower likelihoods of receiving preventive care relative to that of the randomly assigned HMO enrollee.

Average treatment effects of HMO, calculated for a variety of hypothetical individuals, are reported in Table 7. For comparison, we have calculated the effects from our joint model which account for endogeneity of plan-type and from single-equation models which do not account for endogeneity. Because the outcome variables are binary, these treatment effects are the changes in probabilities of receiving the preventive health care services. Once again, the hypothetical individuals we consider have the average characteristics of the entire sample, of black individuals, of non-black individuals and of males and females, of sick and healthy, of those actually enrolled in HMOs and a hypothetical individual with median values of characteristics.

Individuals enrolled in HMO plans (relative to NMC) are 10, 28, 21 and 20 percentage points more likely to receive blood pressure checks, cholesterol exams, flu shots

and mammograms, respectively. The effect of HMO on pap smear tests is not significant. These estimated plan impact effects on probabilities of service are between 2 and 10 times larger as compared to estimates assuming exogeneity of HMO status. Moreover, although there are significant and substantial HMO effects on $flushot$ and $mammogram$ when the endogeneity of health-plan type is considered, these effects are small and insignificant in the single-equation models that do not account for endogeneity. For example, when estimated under exogeneity assumptions, the probability of an "average" individual receiving a blood pressure check increases by 2.6 percentage points. Once self-selection into HMO is taken into account, being enrolled in an HMO increases the probability of a blood pressure check by 10 percentage points. Once again, the average treatment effect accounting for endogeneity is considerably larger.

The effect sizes obtained for the "average individual" are very similar to those obtained for characteristics set at the sample averages of individuals who are actually enrolled in HMOs, i.e., the "average treated individual". For preventive care, however, there is no clear relationship between the treatment effects calculated for the median individual as compared to the treatment effects calculated for the average individual. The effect of HMO enrollment is smaller for the average black individual as compared to the average non-black for bloodpressure, cholesterol and flushot. The average male has a greater HMO effect than an average female with respect to blood pressure and cholesterol checks but the relative effect size is reversed for flu shots. A similar pattern is observed when one compares effects sizes for the healthy as compared to the sick. Generally, the magnitude of the effect across hypothetical females is very similar for pap smear and mammogram.

5.4. Quantifying Selection Effects

Although we have reported on the statistical significance of selection effects, here we quantify their importance in two ways. First, we report the difference between marginal effects of HMO enrollment obtained under exogeneity and the corresponding marginal effects from the simultaneous equations model. If selection effect is significant, then the causal impact of an insurance plan on utilization is not identified under the exogeneity assumption. Consequently, the marginal effect calculated under exogeneity assumptions

incorporates both the causal treatment effect and the selection effect. By contrast, the marginal effect estimated from the correctly specified simultaneous equations model identifies the causal parameter of interest, i.e. the average treatment effect that would be estimated if the data were obtained under an experimental design with randomly assigned treatment. The estimates of the differences, reported in Table 8, are linear approximations to the magnitudes of selection effects. In the case of curative care, the magnitude of the selection effect into HMO plans is similar in magnitude to the treatment effect of HMO enrollment for all types of care except non-physician officed-based care. For example, while a person randomly assigned to an HMO would have 2.6 more visits per year compared to a person assigned to a NMC plan, a person who chose to enroll in an HMO would have 2.4 fewer visits to a doctor per year compared to a person who chose to enroll in a NMC plan, all else equal. For preventive care, the magnitudes of selection and treatment effects are similar in each case. For example, while a person randomly assigned to an HMO would be 28 percentage points more likely to have a cholesterol test compared to a person assigned to a NMC plan, a person who chose to enroll in an HMO would be 22 percentage points less likely to have a cholesterol test compared to a person who chose to enroll in a NMC plan, all else equal. Thus, selection effects are not simply statistically significant, they are substantively important too.

Second, we calculate a dependence measure between the composite errors of the equations that is an analog of the usual Pearson correlation. The composite error for the j^{th} choice in the MMNL is given by $(l_{ji} + \eta_{ji})$. The latent factors l_{ji} are drawn from $N(0, 1)$ and η_{ji} is from a logistic density. The composite error for the outcome equation is $(\sum_j \lambda_j l_{ji} + \varepsilon_i)$ where ε_i is distributed as $\log \Gamma$ in the case of curative outcomes and

$N(0,1)$ in the case of preventive outcomes. Therefore,

$$\begin{aligned} Var(l_{ji}+\eta_{ji}) &= 1+\pi^2/3 \\ Var(\textstyle\sum_j \lambda_j l_{ji}+\varepsilon_i) &= \textstyle\sum_j \lambda_j^2 + \Psi'(1/\alpha) \text{ if outcome is count;} \\ Var(\textstyle\sum_j \lambda_j l_{ji}+\varepsilon_i) &= \textstyle\sum_j \lambda_j^2 + 1 \text{ if outcome is binary;} \\ Cov(l_{ji}+\eta_{ji}, \textstyle\sum_j \lambda_j l_{ji}+\varepsilon_i) &= \lambda_j, \\ Corr(l_{ji}+\eta_{ji}, \textstyle\sum_j \lambda_j l_{ji}+\varepsilon_i) &= \dfrac{Cov(l_{ji}+\eta_{ji}, \sum_j \lambda_j l_{ji}+\varepsilon_i)}{\sqrt{Var(l_{ji}+\eta_{ji}) \times Var(\sum_j \lambda_j l_{ji}+\varepsilon_i)}}, \end{aligned} \tag{13}$$

where Ψ denotes the digamma function and Ψ' is the trigamma function. However, given the assumed factor structure, these correlations have narrower bounds than those for the Pearson correlation. These estimates, also reported in Table 8, are negative and large in most cases, the exceptions being for non-doctor visits and hospital discharges.

5.5. Robustness Checks

Estimates of complex econometric models can be sensitive to choices of samples and covariates, distributional assumptions and parametric functional forms. In order to inform on such issues, our estimated models are subjected to six robustness checks, two involving variations in the sample coverage, two more for the unobserved heterogeneity assumption, and the last two in respect of the estimation method used. These results are summarized in Tables 9 and 10 which report parameter estimates and treatment effects of HMO for curative and preventive care respectively.

Our first robustness check examines the sensitivity of estimated parameters to variations in sample coverage. Recall that our results identify the causal effects of plan type conditional on having insurance coverage. But we have argued that employment status and self-employment status among those employed may be jointly determined with the desire to have access to a particular type of health plan. Therefore, we estimate models for a subsample of those who are employed and a subsample of only those who are employees of firms using only firm characteristics as instruments. Tables 9 and 10 show how the estimated treatment effect of HMO changes if we reestimate our models after excluding first the unemployed and then both the unemployed and the self-employed

from the full sample. For example, the qualitative impact on doctor visits is to reduce the estimate without much change in the standard error. Relative to the full sample, the treatment effect of HMO drops to 2.53 and 2.24 visits compared to 2.65 in the full sample. When the same exercise is carried out for outpatient visits, hospital discharges, emergency room visits, the results regarding the HMO impact are, after allowing for the expected sampling variation, very similar to those for the full sample. For all five preventive measures estimated impact retains the same sign and roughly the same size as in the full original sample.

Our second check involves the impact of using alternative distributional assumptions for the latent factors. In place of normality we assume that the latent factors are drawn from beta distributions centered at zero with unit variance. We consider two cases. In the first, the parameters of the beta density are chosen to have skewness equal to 0.5 and in the second, the selected parameters give a skewness of -0.5. The impact of applying the MSL methodology to these new specifications on the conclusions about the point estimates of the impact of HMO is fairly small. Broadly, the count outcomes show relatively greater sensitivity than the binary outcomes.

Because, our estimation procedure is of full information variety, and such procedures may be sensitive to model misspecification, our third robustness check involves using two simpler "limited information" alternatives based on the instrumental variables method. We first use the linear instrumental variables (LIV) method which is formally applicable to models with linear equations for outcomes and endogenous regressors. For curative utilization the outcome equation linearized by taking logarithms[9] In the case of preventive utilization variables the outcome is a dummy variable, and hence the outcome model is of a linear probability equation, which is a popular choice in the empirical literature. Linearizing the choice probability functions is especially troublesome because of the multinomial nature of the choices. The results in Tables 9 and 10 show that the treatment effect of HMO enrollment on counted outcomes are estimated with considerably less precision in all cases. For three of five outcomes, non-doctor, outpatient and emergency room visits, the magnitudes are much larger than those obtained using

[9] A small positive value is added to the count to avoid definitional problems for zero counts.

MSL methods; in fact they are too large to be plausible. For binary outcomes, the LIV estimates have smaller standard errors so that statistical significance often mirrors those obtained using MSL methods. However, for three of five outcomes, blood pressure and cholesterol checks, and flu shots, the treatment effects obtained using LIV are much bigger than the corresponding effects obtained using MSL methods, and are implausibly large. We also mimic the linear two stage least squares approach by estimating linear probability equations for HMO and OMC choices and then substituting the fitted probabilities in place of the respective dummy variables in the utilization equations. Such procedures are sometimes employed for convenience (Dubin and McFadden, 1984; Johnson and Crystal, 2000) although they may not be consistent. These results are shown against the label "models with fitted plan choice" and are very similar to those obtained using the LIV method, generally imprecise and implausibly large.

Finally, we describe two additional sets of robustness checks. First, a potential source of misspecification is in our choice of functional forms, especially as it relates to covariates. If our specification of covariates is not sufficiently rich, a finding of selection might simply be due to omitted nonlinearities. In preliminary analyses, we explored a variety of quadratic and interaction effects of covariates (including quadratic terms for *age, famsize*, and *income* and interactions of these covariates with gender and minority status). None of these effects was found to be consistently significant across specifications, in part because we have restricted our sample to a relatively homogeneous group. Second, we have estimated models using the sequential limited information maximum simulated likelihood estimator as described in section 3.2. As one might expect given the structure of our latent factors model, the point estimates of the parameters are very close to the estimates from the full-information maximum simulated likelihood approach. The standard errors from the second-stage estimator are, however, 10-20 percent smaller on average. These are incorrect (one expects them to be too small in general) because they do not take estimation uncertainty from the first stage into account. Development of corrected standard errors for the limited information approach is beyond the scope of this paper.

In summary, the results of the robustness exercise provide strong support for the use of a structural latent variable framework to obtain efficient estimates of the key

parameters. There are, however, two alternative specifications we have not implemented that are worthy of mention. First, we have not considered flexible conditional densities for the outcome equations as proposed by Deb and Trivedi (1997) or Gilleskie and Mroz (2004) because incorporating flexible densities and endogenous regressors is a difficult computational problem beyond the scope of this paper, although clearly deserving of further investigation. Both Deb and Trivedi (1997) and Gilleskie and Mroz (2004) model health care use with finite mixtures of densities, albeit in different ways, and find improvement in fit relative to more standard approaches. Neither of these papers considers the issue of endogenous regressors, however. Second, we have not considered a multinomial probit model (MNP) for insurance choices. Instead, we have used the MMNL model with the independence from irrelevant alternatives (IIA) property. It is desirable that we relax this strong assumption that will fail to hold if, for example, the plan choices are not distinct alternatives. The multinomial probit (MNP) model is a leading flexible alternative to the MMNL, but its use in the present context is not feasible because the identification of the covariance structure in the MNP model requires alternative-variant exclusion restrictions. When alternative-specific covariates such as prices are available, as is usually the case in models of transportation choice, the identifying information exists in a usable form. However, here all data are individual-specific and generation of alternative-specific covariates can be done only somewhat arbitrarily (see, for example, Lechner, 2002). Finally, note that even with alternative specific covariates identification of the MNP can be quite fragile (Keane, 1992).

6. Conclusion

This paper shows the feasibility of estimating nonlinear simultaneous equations models for discrete outcomes, with a large number of unknown coefficients, using computer intensive methods. When this methodology is applied to model the choice of health insurance plans and health care utilization using MEPS data for the US nonelderly population, we find significant evidence of selection bias. This contradicts the implicit assumption of negligible selection bias in existing econometric research on health care utilization that has assumed exogeneity of insurance plans. Taking the endogeneity of

insurance status into account has substantial impact on the magnitudes of insurance effects.

Data limitations prevent us from addressing a number of important issues. Labels such as HMO and OMC are ultimately meaningful only if they adequately capture the underlying restrictions placed on consumer access to care, and the underlying cost sharing features. Replacing the broad labels by variables that reflect the economic attributes of health plans is desirable if we wish to explain why there is selection into managed care organizations and whether there might be differential quality and ease of use issues in utilization of care (Kemper, et al., 2002). Second, because we do not have information on prices or copayments associated with alternative health plans, we cannot evaluate the impact of cost-sharing. Finally, MEPS is a short panel with $T = 2$ at most, so we cannot adequately address dynamic issues. It is possible to modify our econometric model, however, to address such questions if appropriate data were available.

Applied econometrics contains many variants of the non-normal selection model, some of which involve the appearance of endogenous treatment dummy variables in equations with discrete and/or censored outcomes. The approach developed here can be extended to these cases even if treatments are truncated, censored or continuous, instead of binary or multinomial. However, because our computational methods are very time intensive, further research is needed to investigate other promising computational methodologies that would efficiently handle larger models and samples than those used in this article.

References

Bhat, C.R. (2001) "Quasi-random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model", *Transportation Research: Part B,* 35, 677-693.

Bingley, Paul; Walker, Ian (2001), "Household Unemployment and the Labor Supply of Married Women", *Economica*, 68, 157-85.

Brownstone, D. and K. Train (1999), "Forecasting New Product Penetration with Flexible Substitution Patterns", *Journal of Econometrics*, 89, 109-129.

Cao, Z. and T.G. McGuire (2002), "Service-Level Selection by HMOs in Medicare", *Working Paper.*

Cardon, J.H. and I. Hendel (2001) "Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey", *RAND Journal of Economics*, 32, 3, 408-27.

Carrasco, R. (2001), "Binary Choice with Binary Endogenous Regressors in Panel Data: Estimating the Effect of Fertility on Female Labor Participation", *Journal of Business and Economic Statistics*, 19, 385-394.

Deb, P., and P.K. Trivedi (1997), "Demand for Medical Care by the Elderly in the United States: A Finite Mixture Approach", *Journal of Applied Econometrics*, 12, 313-336.

Dowd, B., R. Feldman, S. Cassou, and M. Finch (1991), "Health Plan Choice and the Utilization of Health Care Services", *Review of Economics and Statistics*, 73, 85-93.

Dranove, D. (2000), *The Economic Evolution of American Health Care: From Marcus Welby to Managed Care,* Princeton: Princeton University Press.

Dubin, J.A. and D. McFadden (1984), "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption", *Econometrica*, 52, 345-362.

Ettner, S.L. (1997), "Adverse Selection and the Purchase of Medigap Insurance by the Elderly", *Journal of Health Economics*, 16, 543-62.

Frank, R.G., J. Glazer and T.G. McGuire (2000), "Measuring Adverse Selection in Managed Health Care", *Journal of Health Economics*, 19, 829-854.

Geweke, J., G. Gowrisankaran, R.J. Town (2003), "Bayesian Inference for Hospital Quality in a Selection Model", *Econometrica*, 71(4), 1215-38.

Gilleskie, D.B. and T.A. Mroz (2004), "A Flexible Approach for Estimating the Effects of Covariates on Health Expenditures", *Journal of Health Economics*, 23, 391-418.

Glied, S. (2000), "Managed Care", Chapter 13 in Culyer, A.J. and J.P. Newhouse, Editors, *Handbook of Health Economics,* Vol 1A, 707-745.

Goldman, D.P. (1995), "Managed Care as a Public Cost-Containment Mechanism", *Rand Journal of Economics*, 26, 277-295.

Goldman, D.P., S.D. Hosek, L.S. Dixon, E.M. Sloss (1995), "The Effects of Benefit Design and Managed Care on Health Care Costs", *Journal of Health Economics*, 14, 401-418.

Gouriéroux, C. and A. Monfort (1996), *Simulation Based Econometrics Methods*, New York: Oxford University Press.

Gowrisankaran, G., and Town, R. J, (1999) "Estimating the Quality of Care in Hospitals Using Instrumental Variables", *Journal of Health Economics*, 18, 747-767.

Gruber, J (2000), "Tax Subsidies for Health Insurance: Evaluating the Costs and Benefits", National Bureau of Economic Research, Inc, NBER Working Papers: 7553.

Heckman, J.J., J. Smith and N. Clements (1997), "Making the Most of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts", *The Review of Economic Studies*, 64, 487-535.

Heckman, J.J., and E. Vytlacil (2001), "Policy-Relevant Treatment Effects", *American Economic Review Papers and Proceedings,* 91, 107-111.

Hillman, A., M. Pauly, and J. Kerstein (1989), "How Do Financial Incentives Affect Physicians' Clinical Decisions and Financial Performance of Health Maintenance Organizations?", *New England Journal of Medicine*, 321, 86-92.

Hyslop, D.R. (1999), "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women", *Econometrica*, 67, 1255-1294.

Imbens, G.W. and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, 467-475.

Jensen, Eric R, (1999), "An Econometric Analysis of the Old-Age Security Motive for Childbearing", *International Economic Review*, 31, 953-968.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, London: Chapman and Hall.

Johnson, R.W. and S. Crystal (2000), "Uninsured Status and Out-of-pocket Costs at Midlife", *Health Services Research*, 35, 911-932.

Keane, M.P. (1992), "A Note on Identification in the Multinomial Probit Model", *Journal of Business and Economic Statistics,* 10, 193-200.

Kemper, P., H. T. Tu, J. D. Reschovsky, E. Schaefer (2002) "Insurance Product Design and Its Effects: Trade-Offs Along the Managed Care Continuum", *Inquiry*, 39, 101-117.

Kenkel, D.S. and J. Terza (2001), "The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect", *Journal of Applied Econometrics*, 16, 165-84.

Lee, L-F. (1983), "Generalized Econometric Models with Selectivity", *Econometrica,* 51, 507-512.

Lechner, M. (2002) "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies", *Review of Economics and Statistics*, 84, 205-220.

Lindsay, B. J. (1995), *Mixture Models: Theory, Geometry, and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, IMS-ASA.

McFadden, D. (1980) "Econometric Models for Probabilistic Choice among Products", *Journal of Business,* 53, S13-S29.

McFadden, D. and K. Train (2000), "Mixed MNL Models for Discrete Response", *Journal of Applied Econometrics*, 15, 447-470.

Meer, J. and H.S. Rosen (2003), "Insurance and the Utilization of Medical Services", NBER Working Paper 9812.

Mello, M.M., S.C. Stearns, and E.C. Norton (2002), "Do Medicare HMOs Still Reduce Health Services Use After Controlling for Selection Bias?", *Health Economics,* 11, 323-340.

Miller, R.H. and H.S. Luft (1994), "Managed Care Plan Performance Since 1980", *Journal of American Medical Association*, 271, 1512-1519.

Mroz, T.A. (1999), "Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome", *Journal of Econometrics*, 92, 233-274.

Olsen, R. and G. Farkas (1989), "Endogenous Covariates in Duration Models and the Effect of Adolescent Childbirth on Schooling", *Journal of Human Resources*, 24(1), 39-53.

Pagan, A.R. (1986), "Two Stage and Related Estimators and Their Properties", *Review of Economic Studies,* 53, 517-538.

Pitt, M.R. and M. Rosenzweig (1990) "Estimating the Intrahousehold Incidence of Illness: Child Health and Gender-Inequality in the Allocation of Time", *International Economic Review*, 31(4), 969-80

Prieger, J. (2002), "A Flexible Parametric Selection Model for Non-Normal Data with Application to Health Care Usage", *Journal of Applied Econometrics,* 17, 367-392.

Terza, J.V. (1998), "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects", *Journal of Econometrics*, 84, 129-154.

Train, K. (2002). *Discrete Choice Methods with Simulation,* New York: Cambridge University Press.

Van Ophem, H. (2000), "Modeling Selectivity in Count-Data Models", *Journal of Business and Economic Statistics*, 18, 503-511.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica,* 50, 1-25.

Table 1: Descriptive Statistics for Utilization

Variable	Definition	N	HMO	OMC	NMC
Curative utilization			47%	8.2%	44.8%
Doctor	number of visits to a physician in an office setting	8129	3.279	3.499*	3.066
Nondoctor	number of non-physician visits in an office setting	8129	1.378	1.694	1.423
Outpatient	number of hospital outpatient visits	8129	0.386*	0.368	0.545
Hospital	number of hospital discharges	8129	0.076	0.070	0.082
ER	number of emergency room visits	8129	0.141	0.124	0.138
Preventive utilization					
Bloodpressure	=1 if blood pressure was checked in last two years	7952	0.923*	0.924*	0.887
Cholesterol	=1 if cholesterol was checked in last two years	7717	0.630*	0.595	0.564
Flushot	=1 if flu shot was received in the last year	7948	0.198	0.185	0.198
Papsmear	=1 if pap smear test was received in the last year	4082	0.679*	0.694*	0.617
Mammogram	=1 if mammogram was received in the last year	2105	0.537	0.531	0.521

Note: * indicates that the estimate is significantly different from the base case (NMC) at the 5 percent level.

Table 2: Descriptive Statistics for Explanatory Variables

Variable	Definition	HMO	OMC	NMC
Demographic characteristics				
familysize	family size	3.080	2.993	3.042
age	age/10	3.944*	3.937*	4.091
education	years of school	13.474	13.572	13.395
income	income/1000	39.014*	37.514	37.272
female	=1 if female	0.534*	0.513	0.505
black	=1 if black	0.122*	0.118*	0.084
hispanic	=1 if hispanic	0.158*	0.125*	0.093
married	=1 if married	0.674	0.715	0.683
northeast	=1 if north east	0.212*	0.183	0.194
midwest	=1 if midwest	0.200*	0.244*	0.295
south	=1 if south	0.321*	0.367	0.352
msa	=1 if metropolitan statistical area	0.871*	0.906*	0.697
Employment characteristics				
employed	=1 if employed	0.888*	0.887	0.863
selfemployed	=1 if self employed	0.083*	0.082*	0.130
firmsize	firm size/10	14.675*	14.481*	10.581
multlocation	=1 if multiple locations	0.594*	0.629*	0.499
union	=1 if union	0.148*	0.124	0.129
govtjob	=1 if government job	0.183*	0.151	0.161
blue	=1 if blue coller	0.223	0.204	0.216
service	=1 if service	0.356	0.392*	0.346
miscellaneous	=1 if miscellaneous industry	0.086	0.086	0.076
Health status				
verygood	=1 if very good health	0.356	0.349	0.365
good	=1 if good health	0.239	0.249	0.227
fair	=1 if fair health	0.061	0.052	0.056
poor	=1 if poor health	0.012	0.010	0.014
chronic	number of chronic conditions	0.541	0.517	0.535
physicallim	=1 if physical limitation	0.056	0.064	0.059

Note: * indicates that the estimate is significantly different from the base case (NMC) at the 5 percent level.

Table 3: Marginal Effects in MMNL Insurance Plan Choice Model

	Pr(NMC)		Pr(OMC)		Pr(HMO)	
Variable	Marg.	St. err.	Marg.	St. err.	Marg.	St. err.
familysize	0.009*	0.004	-0.006*	0.002	-0.003	0.004
age	0.034*	0.006	-0.006*	0.003	-0.028*	0.006
married	-0.050*	0.015	0.029*	0.007	0.021	0.014
northeast	0.102*	0.018	-0.009	0.009	-0.093*	0.017
midwest	0.175*	0.018	0.002	0.010	-0.177*	0.016
south	0.127*	0.016	0.006	0.009	-0.133*	0.015
msa	-0.241*	0.014	0.057*	0.006	0.185*	0.014
income	-3e-4	2e-4	-2e-4	1e-4	0.001*	2e-4
female	-0.035*	0.012	-0.005	0.006	0.040*	0.012
black	-0.085*	0.018	0.006	0.011	0.079*	0.019
hispanic	-0.082*	0.018	-0.004	0.009	0.085*	0.018
education	4e-4	0.003	0.000	0.002	-0.001	0.003
employed	0.014	0.022	-0.001	0.013	-0.013	0.023
selfemployed	0.050*	0.023	-1e-4	0.013	-0.050*	0.022
firmsize	-0.002*	4e-4	3e-4	2e-4	0.002*	4e-4
multlocation	-0.057*	0.015	0.024*	0.008	0.033*	0.015
union	0.010	0.018	-0.012	0.009	0.002	0.017
govtjob	-0.015	0.019	-0.016	0.009	0.031	0.019
blue	-0.009	0.019	-0.008	0.010	0.017	0.019
service	-2e-4	0.016	0.002	0.009	-0.001	0.015
physicallim	-0.024	0.026	0.017	0.016	0.007	0.027
chronic	-0.009	0.008	-0.001	0.004	0.010	0.008
verygood	-0.002	0.015	-0.004	0.007	0.006	0.014
good	-0.033*	0.016	0.007	0.008	0.026	0.016
fair	-0.020	0.027	-0.007	0.014	0.027	0.028
poor	0.026	0.053	-0.014	0.027	-0.012	0.052

Note: * indicates that the parameter estimate is significantly different from zero at the 5 percent level.

Table 4: Insurance and factor loading parameters: curative health care services

	Doctor		Nondoctor		Outpatient		Hospital		ER	
	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.
OMC	0.126	0.158	0.216	0.158	0.635*	0.173	0.378	0.568	0.296	0.325
HMO	0.906*	0.051	0.047	0.086	1.396*	0.107	-0.547	0.455	0.928*	0.149
λ_{OMC}	0.136	0.169	0.001	0.001	-0.878*	0.088	-0.460	0.569	-0.348	0.325
λ_{HMO}	-0.934*	0.047	-0.001	0.002	-1.686*	0.078	0.648	0.532	-1.004*	0.156

Note: * indicates that the parameter estimate is significantly different from zero at the 5 percent level.

Table 5: Average treatment effects of HMO: curative health care services

	Doctor		Nondoctor		Outpatient		Hospital		ER	
	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.
	accounting for endogeneity of health plan choice									
mean	2.649*	0.197	0.047	0.085	0.531*	0.075	-0.035	0.096	0.127*	0.031
median	1.898*	0.151	0.038	0.070	0.222*	0.041	-0.033	0.105	0.056*	0.016
black	1.895*	0.164	0.041	0.062	0.277*	0.087	-0.021	0.094	0.114*	0.031
non black	3.065*	0.220	0.105	0.101	0.679*	0.090	-0.034	0.091	0.154*	0.031
male	1.955*	0.142	0.063	0.060	0.405*	0.058	-0.025	0.066	0.156*	0.031
female	3.905*	0.279	0.119	0.119	0.771*	0.103	-0.045	0.136	0.150*	0.031
chronic>0	4.998*	0.364	0.185	0.186	1.137*	0.159	-0.048	0.140	0.197*	0.040
chronic=0	1.952*	0.137	0.056	0.053	0.367*	0.049	-0.027	0.076	0.131*	0.026
in HMO's	2.812*	0.179	0.088	0.083	0.492*	0.063	-0.035	0.108	0.148*	0.030
	assuming exogeneity of health plan choice									
	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.
mean	0.247*	0.090	0.043	0.083	-0.035	0.044	-0.004	0.006	0.013	0.009
median	0.194*	0.070	0.034	0.067	-0.020	0.027	-0.004	0.006	0.007	0.005
black	0.194*	0.070	0.031	0.061	-0.040	0.060	-0.003	0.005	0.014	0.009
non black	0.254*	0.093	0.045	0.087	-0.034	0.043	-0.004	0.006	0.013	0.009
male	0.174*	0.064	0.030	0.060	-0.029	0.037	-0.003	0.004	0.013	0.009
female	0.341*	0.124	0.059	0.114	-0.042	0.053	-0.005	0.008	0.014	0.009
chronic>0	0.433*	0.157	0.093	0.181	-0.066	0.084	-0.006	0.009	0.018	0.012
chronic=0	0.174*	0.064	0.026	0.051	-0.023	0.030	-0.003	0.005	0.011	0.008
in HMO's	0.246*	0.089	0.042	0.081	-0.034	0.044	-0.004	0.006	0.013	0.009

Note: * indicates that the parameter estimate is significantly different from zero at the 5 percent level.

Table 6: Insurance and factor loading parameters: preventive health care services

	Bloodpressure		Cholesterol		Flushot		Papsmear		Mammogram	
	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.
OMC	0.365	0.471	0.234	0.321	0.206	0.377	-0.282	0.385	-1.129*	0.482
HMO	1.032*	0.403	1.141*	0.277	1.532*	0.209	0.610	0.540	1.050	0.622
λ_{OMC}	-0.077	0.502	-0.148	0.341	-0.193	0.392	0.508	0.424	1.388*	0.447
λ_{HMO}	-0.927*	0.430	-1.120*	0.302	-1.750*	0.229	-0.570	0.615	-1.177	0.727

Note: * indicates that the parameter estimate is significantly different from zero at the 5 percent level.

Table 7: Average treatment effects of HMO: preventive health care services

	Bloodpressure		Cholesterol		Flushot		Papsmear		Mammogram	
	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.
	accounting for endogeneity of health plan choice									
mean	0.102*	0.026	0.283*	0.035	0.210*	0.015	0.177	0.119	0.199*	0.091
median	0.139*	0.036	0.294*	0.036	0.195*	0.016	0.167	0.116	0.199*	0.091
black	0.076*	0.029	0.260*	0.033	0.185*	0.014	0.280*	0.116	0.241*	0.090
non black	0.093*	0.026	0.296*	0.035	0.213*	0.015	0.281*	0.119	0.241*	0.091
male	0.123*	0.036	0.297*	0.035	0.204*	0.015	–	–	–	–
female	0.065*	0.019	0.289*	0.034	0.217*	0.015	–	–	–	–
chronic>0	0.048*	0.016	0.256*	0.030	0.243*	0.017	0.279*	0.118	0.240*	0.090
chronic=0	0.124*	0.036	0.303*	0.036	0.188*	0.014	0.283*	0.120	0.241*	0.091
in HMO's	0.090*	0.027	0.294*	0.035	0.204*	0.015	0.281*	0.118	0.241*	0.091
	assuming exogeneity of health plan choice									
	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.	Marg.	St. Err.
mean	0.026*	0.005	0.058*	0.012	0.013	0.009	0.045*	0.016	0.023	0.024
median	0.033*	0.008	0.061*	0.013	0.013	0.009	0.042*	0.015	0.023	0.024
black	0.024*	0.006	0.051*	0.011	0.011	0.008	0.043*	0.015	0.022	0.024
non black	0.026*	0.005	0.059*	0.012	0.013	0.009	0.045*	0.016	0.023	0.024
male	0.040*	0.008	0.060*	0.012	0.012	0.009	–	–	–	–
female	0.016*	0.003	0.057*	0.012	0.014	0.010	–	–	–	–
chronic>0	0.011*	0.003	0.050*	0.010	0.016	0.011	0.044*	0.016	0.023	0.024
chronic=0	0.039*	0.008	0.061*	0.013	0.011	0.008	0.045*	0.016	0.023	0.024
in HMO's	0.026*	0.006	0.058*	0.012	0.013	0.009	0.045*	0.016	0.023	0.024

Note: * indicates that the parameter estimate is significantly different from zero at the 5 percent level.

Table 8: Magnitudes of Selection Effects into HMO's

Variable	Treatment effect	Selection effect	Error correlation
Curative utilization			
Doctor	2.649	-2.402	-0.410
Nondoctor	0.047	-0.004	-0.0001
Outpatient	0.531	-0.566	-0.344
Hospital	-0.035	0.031	0.150
ER	0.127	-0.114	-0.352
Preventive utilization			
Bloodpressure	0.102	-0.076	-0.328
Cholesterol	0.283	-0.225	-0.358
Flushot	0.210	-0.197	-0.417
Papsmear	0.177	-0.132	-0.219
Mammogram	0.199	-0.176	-0.274

Table 9: HMO Effects on Curative Care: Alternative Models and Samples

Model	N	Coeff.	St. Err.	Marg.	St. Err.
			doctor visits		
Sample without unemployed	7127	0.914*	0.061	2.531*	0.293
Sample without unemployed and self employed	6285	0.803*	0.065	2.246*	0.201
Beta distributed latent factors skewness = 0.5	8129	0.906*	0.055	2.649*	0.212
Beta distributed latent factors skewness = -0.5	8129	0.906*	0.052	2.682*	0.202
Linear instrumental variables	8129	1.038	1.284	1.038	1.284
Models with fitted insurance plan choice	8129	0.549	0.325	1.414	0.838
			non doctor visits		
Sample without unemployed	7127	0.043	0.094	0.040	0.090
Sample without unemployed and self employed	6285	0.036	0.100	0.035	0.092
Beta distributed latent factors skewness = 0.5	8129	0.047	0.086	0.047	0.085
Beta distributed latent factors skewness = -0.5	8129	0.047	0.086	0.047	0.085
Linear instrumental variables	8129	2.653	1.731	2.653	1.731
Models with fitted insurance plan choice	8129	1.086	0.733	1.072	0.724
			outpatient visits		
Sample without unemployed	7127	1.411*	0.126	0.556*	0.098
Sample without unemployed and self employed	6285	1.397*	0.128	0.526*	0.089
Beta distributed latent factors skewness = 0.5	8129	1.457*	0.117	0.598*	0.103
Beta distributed latent factors skewness = -0.5	8129	1.385*	0.107	0.628*	0.099
Linear instrumental variables	8129	0.950	0.789	0.95	0.789
Models with fitted insurance plan choice	8129	2.681*	0.908	0.816*	0.277
			hospital discharges		
Sample without unemployed	7127	-0.552*	0.210	-0.032*	0.015
Sample without unemployed and self employed	6285	-0.532*	0.262	-0.032	0.021
Beta distributed latent factors skewness = 0.5	8129	-0.511	0.384	-0.032	0.059
Beta distributed latent factors skewness = -0.5	8129	-0.544*	0.267	-0.035	0.022
Linear instrumental variables	8129	-0.027	0.090	-0.027	0.090
Models with fitted insurance plan choice	8129	-0.531	1.133	-0.032	0.069
			emergency room visits		
Sample without unemployed	7127	0.935*	0.162	0.127*	0.032
Sample without unemployed and self employed	6285	0.928*	0.241	0.135	0.119
Beta distributed latent factors skewness = 0.5	8129	0.738*	0.166	0.097*	0.027
Beta distributed latent factors skewness = -0.5	8129	0.726*	0.153	0.096*	0.026
Linear instrumental variables	8129	0.205	0.126	0.205	0.126
Models with fitted insurance plan choice	8129	1.421	0.854	0.170	0.102

Note: * indicates that the parameter estimate is significantly different from zero at the 5 percent level.

Table 10: HMO Effects on Preventive Care: Alternative Models and Samples

Model	N	Coeff.	St. Err.	Marg.	St. Err.
			blood pressure check		
Sample without unemployed	6969	1.034	0.592	0.104*	0.049
Sample without unemployed and self employed	6137	1.045*	0.318	0.107*	0.019
Beta distributed latent factors skewness = 0.5	7952	2.131*	1.066	0.135*	0.040
Beta distributed latent factors skewness = -0.5	7952	1.029*	0.377	0.102*	0.024
Linear instrumental variables	7952	0.138	0.077	0.180*	0.079
Models with fitted insurance plan choice	7952	1.397*	0.490	0.174*	0.063
			cholesterol check		
Sample without unemployed	6763	1.173*	0.220	0.289*	0.026
Sample without unemployed and self employed	5959	1.107*	0.210	0.279*	0.027
Beta distributed latent factors skewness = 0.5	7717	1.173*	0.309	0.285*	0.041
Beta distributed latent factors skewness = -0.5	7717	1.162*	0.291	0.285*	0.036
Linear instrumental variables	7717	0.494*	0.137	0.494*	0.137
Models with fitted insurance plan choice	7717	1.507*	0.356	0.577*	0.136
			flu shot		
Sample without unemployed	6971	1.557*	0.192	0.210*	0.013
Sample without unemployed and self employed	6145	1.591*	0.212	0.211*	0.013
Beta distributed latent factors skewness = 0.5	7948	1.532*	0.234	0.214*	0.016
Beta distributed latent factors skewness = -0.5	7948	1.530*	0.220	0.213*	0.015
Linear instrumental variables	7948	0.539*	0.126	0.539*	0.126
Models with fitted insurance plan choice	7948	2.084*	0.386	0.542*	0.100
			mammogram		
Sample without unemployed	1675	0.937	0.817	0.192	0.119
Sample without unemployed and self employed	1490	1.427*	0.668	0.273*	0.081
Beta distributed latent factors skewness = 0.5	2105	1.032	0.803	0.197	0.113
Beta distributed latent factors skewness = -0.5	2105	1.033	0.602	0.196*	0.089
Linear instrumental variables	2105	0.149	0.161	0.149	0.161
Models with fitted insurance plan choice	2105	0.537	0.415	0.214	0.165
			papsmear		
Sample without unemployed	3357	0.613	0.780	0.178	0.140
Sample without unemployed and self employed	3040	0.647	0.835	0.188	0.159
Beta distributed latent factors skewness = 0.5	4082	0.609	0.610	0.175	0.130
Beta distributed latent factors skewness = -0.5	4082	0.609	0.629	0.175	0.128
Linear instrumental variables	4082	0.136	0.153	0.136	0.153
Models with fitted insurance plan choice	4082	0.355	0.427	0.131	0.157

Note: * indicates that the parameter estimate is significantly different from zero at the 5 percent level.

Appendix A
Simulation Acceleration

Estimation by MSL requires computer-generated draws of random numbers. Typically, pseudo-random numbers are used. In univariate cases, a small number of pseudo-random draws is sufficient to reduce the simulation error to acceptable levels. However, many more draws are required in multidimensional cases to achieve a similar level of accuracy (Brownstone and Train, 1999; Train, 2002). In addition, our limited experimental evidence shows that many more simulation draws are needed in simultaneous equation systems as compared to systems with correlated errors, but without endogenous regressors, to achieve similar levels of accuracy. The published literature directly investigating this issue is scanty, but Hyslop (1999) describes similar findings in a different context. This feature makes the estimation of nonlinear simultaneous equations models with large numbers of observations and parameters by MSL computationally infeasible without the application of acceleration techniques to reduce the number of required simulation draws.

Increasing the number of simulation draws to reduce simulation error to acceptable levels is simple in principle but computationally costly. In our case, computational times were prohibitively high when sufficient numbers of pseudo-random draws were used. We therefore draw on the recent adavances in numerical analysis that use intelligent, systematic draws rather than random draws to speed up convergence of the required expectations. The quasi-Monte Carlo method, instead of using S pseudo-random points, makes draws based on a non-random selection of points within the domain of integration. The use of Halton sequences is one such quasi-Monte Carlo method introduced by Bhat (2001) in the context of simulation-based estimation of mixed multinomial models. Halton sequences have two desirable properties vis-a-vis pseudo-random draws. First, they are designed to give more even coverage over the domain of the mixing distribution. With more evenly spread draws for each observation, the simulated probabilities vary less over observations, relative to those calculated with random draws. Second, with Halton sequences, the draws for one observation tend to fill in the spaces left empty by the previous observations. The simulated probabilities are, therefore, negatively correlated over observations. This negative correlation reduces the variance in the simulated likelihood function. Under suitable regularity conditions, the integration error using pseudo-random sequences is in the order of N^{-1} as compared to pseudo-random sequences where the convergence rate is $N^{-1/2}$ (Bhat, 2001).

Halton sequences are best described by example. Consider the prime number 2. Its Halton sequence is constructed as follows. Divide the unit interval (0,1) into 2 parts. The dividing point 1/2 becomes the first element of the Halton sequence. Next divide each part into two more parts. The dividing points, 1/4 and 3/4 become the next two elements of the sequence. Divide each of the four parts into two parts each, and

continue. Halton sequences on non-prime numbers are not unique because the Halton sequence for a non-prime number divides the unit space in the same way as each of the prime numbers that constitute the non-prime. In our model, we have two unobserved factors that need to be integrated out, so we generate two Halton sequences, based on the primes 2 and 3:

$$\begin{array}{cccccc} \{1/2 & 1/4 & 3/4 & 1/8 & 3/8 & ...\} \\ \{1/3 & 2/3 & 1/9 & 2/9 & 4/9 & ...\} \end{array}$$

The length of each sequence is determined by the number of observations N and the numbers of simulation draws S. We discard the first 20 elements of the sequence as the early elements have a tendency to be correlated over Halton sequences with different primes (see Train, 1999, for an example). Consequently, we begin by generating Halton sequences of length $N \times S + 20$ and discard the first twenty elements of each sequence. For each element of each sequence, we calculate the inverse of the cumulative normal distribution. The resulting values are the Halton draws from the mixing distribution. The first group of S elements in the resulting sequence is assigned to the first observation in the sample, the next S elements to the second observation, and so on.

Bhat (2001) and Train (2002) demonstrate dramatic improvement in simulation errors from the use of Halton-sequence based draws relative to the usual pseudo-random draws. Bhat (2001) finds that the simulation error in the estimated parameters was lower using 100 Halton numbers than 1000 random numbers. Train (2002) finds that the variance over draws in the simulated probability for an observation is half as large with 100 Halton draws than 1000 random draws. Our experience in the context of the model considered here suggests less dramatic improvement over random sequences, but the improvement is substantial nevertheless.

Appendix B
Parameter Estimates of Outcome Equations

Table 1: Curative Care

	Doctor		Nondoctor		Outpatient		Hospital		ER	
Variable	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.
intercept	-1.726	0.147	-2.178	0.338	-5.881	0.391	-3.605	0.450	-1.816	0.342
familysize	-0.049	0.012	-0.099	0.032	-0.122	0.035	0.059	0.035	-0.008	0.030
age	0.061	0.015	0.07	0.040	0.268	0.041	-0.033	0.052	-0.184	0.039
education	0.057	0.007	0.097	0.018	0.041	0.019	-0.005	0.021	-0.046	0.017
income	0.001	0.001	0.002	0.001	-0.002	0.001	-2e-4	0.002	8e-5	0.002
female	0.615	0.032	0.561	0.090	0.536	0.090	0.581	0.105	-0.098	0.077
black	-0.353	0.052	-0.376	0.213	-0.37	0.188	-0.157	0.176	-0.103	0.125
hispanic	-0.163	0.052	-0.533	0.136	-0.402	0.153	0.162	0.15	0.006	0.121
married	0.112	0.038	0.037	0.098	0.282	0.108	0.187	0.122	-0.193	0.090
northeast	0.271	0.048	-0.08	0.119	0.882	0.138	0.129	0.152	0.223	0.121
midwest	0.173	0.047	-0.137	0.114	1.033	0.139	-0.129	0.160	0.522	0.116
south	0.209	0.044	-0.204	0.119	0.26	0.136	-0.041	0.139	0.215	0.111
msa	-0.051	0.042	0.019	0.106	-0.467	0.104	0.003	0.178	-0.389	0.094
verygood	0.192	0.038	0.283	0.104	0.292	0.114	-0.003	0.129	0.189	0.097
good	0.416	0.044	0.453	0.122	0.68	0.119	0.457	0.135	0.409	0.104
fair	0.797	0.066	0.349	0.177	1.318	0.178	1.245	0.172	1.089	0.142
poor	0.991	0.130	1.402	0.319	1.811	0.338	2.112	0.268	1.421	0.237
physicallim	0.260	0.062	1.164	0.249	0.486	0.169	0.319	0.170	-0.017	0.143
chronic	0.490	0.019	0.641	0.053	0.463	0.050	0.264	0.057	0.277	0.046
OMC	0.126	0.158	0.216	0.158	0.635	0.173	0.378	0.568	0.296	0.325
HMO	0.906	0.051	0.047	0.086	1.396	0.107	-0.547	0.455	0.928	0.149
α	0.278	0.060	6.163	0.223	1.139	0.199	1.692	0.877	0.574	0.283
λ_{OMC}	0.136	0.169	0.001	0.001	-0.878	0.088	-0.46	0.569	-0.348	0.325
λ_{HMO}	-0.934	0.047	-0.001	0.002	-1.686	0.078	0.648	0.532	-1.004	0.156
log likelihood	-24331		-16668		-12133		-9165		-10322	

Table 2: Preventive Care

	Bloodpressure		Cholesterol		Flushot		Papsmear		Mammogram	
Variable	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.
intercept	-0.733	0.309	-3.837	0.575	-5.037	0.492	-0.272	0.361	-5.386	1.347
familysize	-0.082	0.026	-0.031	0.018	-0.110	0.030	-0.122	0.036	-0.117	0.061
age	0.039	0.030	0.432	0.063	0.463	0.051	-0.092	0.030	0.685	0.169
education	0.085	0.020	0.070	0.013	0.070	0.015	0.060	0.020	0.079	0.029
income	0.002	0.001	0.002	0.001	0.001	0.001	0.002	0.001	0.003	0.002
female	0.729	0.141	0.179	0.052	0.226	0.071	–	–	–	–
black	0.042	0.101	0.446	0.093	-0.454	0.128	0.192	0.091	0.211	0.197
hispanic	-0.077	0.094	0.180	0.074	-0.257	0.116	-0.032	0.097	-0.087	0.216
married	0.237	0.083	0.266	0.066	0.099	0.086	0.411	0.124	0.495	0.188
northeast	0.277	0.111	0.506	0.108	-0.080	0.103	-0.035	0.084	0.444	0.214
midwest	0.171	0.107	0.180	0.081	0.061	0.101	-0.031	0.094	0.306	0.212
south	0.193	0.100	0.367	0.089	0.102	0.095	0.033	0.088	0.106	0.172
verygood	0.120	0.068	0.079	0.055	0.067	0.082	-0.113	0.070	0.020	0.147
good	0.324	0.096	0.176	0.066	0.225	0.091	-0.035	0.075	0.092	0.161
fair	0.557	0.186	0.364	0.120	0.391	0.148	-0.166	0.127	0.302	0.255
poor	0.761	0.417	0.081	0.218	0.128	0.311	-0.382	0.245	0.450	0.485
physicallim	0.140	0.175	0.012	0.105	0.059	0.139	-0.198	0.118	-0.247	0.216
msa	-0.122	0.090	0.154	0.059	-0.396	0.100	0.117	0.098	0.280	0.168
chronic	0.579	0.129	0.377	0.060	0.246	0.047	0.097	0.039	0.083	0.071
OMC	0.365	0.471	0.234	0.321	0.206	0.377	-0.282	0.385	-1.129	0.482
HMO	1.032	0.403	1.141	0.277	1.532	0.209	0.610	0.540	1.050	0.622
λ_{OMC}	-0.077	0.502	-0.148	0.341	-0.193	0.392	0.508	0.424	1.388	0.447
λ_{HMO}	-0.927	0.430	-1.120	0.302	-1.750	0.229	-0.570	0.615	-1.177	0.727
log likelihood	-9085		-11295		-10571		-6085		-3174	