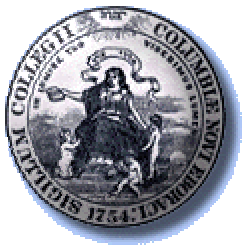

Columbia University

Department of Economics
Discussion Paper Series

MEASURING THE RELATIVE PERFORMANCE OF PROVIDERS OF A HEALTH SERVICE

Daniel A. Ackerberg
Matilde P. Machado
Michael H. Riordan

Discussion Paper #:0102-12

Department of Economics
Columbia University
New York, NY 10027

February 2002

Columbia University
Department of Economics Discussion Paper No. 0102-12
Measuring the Relative Performance of Providers of a Health Service[*†]
Daniel A. Ackerberg
Matilde P. Machado
Michael Riordan
February 2002
JEL No. L22, I11, L11

Abstract:

A methodology is developed and applied to compare the performance of publicly funded agencies providing treatment for alcohol abuse in Maine. The methodology estimates a Wiener process that determines the duration of completed treatments, while allowing for agency differences in the effectiveness of treatment, standards for completion of treatment, patient attrition, and the characteristics of patient populations. Notably, the Wiener process model separately identifies agency fixed effects that describe differences in the effectiveness of treatment ('treatment effects'), and effects that describe differences in the unobservable characteristics of patients ('population effects'). The estimated model enables hypothetical comparisons of how different agencies would treat the same populations. The policy experiment of transferring the treatment practices of more cost effective agencies suggests that Maine could have significantly reduced treatment costs without compromising health outcomes by identifying and transferring best practices.

[†] Research support was provided to Riordan by a grant from the National Institute of Drug Abuse (#1-RO1-DA08715-01) and to Machado by a Marie Curie TMR grant from the European Community (#ERBFMBICT972472) and by a Spanish Education and Culture Ministry grant (#PB98-0137). Part of this research was done while Machado was a postdoctoral fellow at the IAE, and while Riordan was on sabbatical at Yale University. Useful comments and discussions at various stages of this research came from Steven Berry, Zvi Eckstein, Richard Frank, David Genesove, Igal Hendel, Jereal Holly, Thomas McGuire, and Ariel Pakes and from seminar participants at Boston University, Federal Reserve Bank of New York, Yale University, 2000 Latin American and Caribbean Economic Association conference, Fundacion Empresa Publica, Jornadas de Economia Industrial, and the 2000 ASSET Conference, and participants at the NBER Productivity Workshop in December 1998. Deepak Agrawal, Mingshan Lu, and Seth Seabury providedresearch assistance at different stages of the project, and Margaret Commons provided crucial assistance with the data. Mercedes Cabanas provided library assistance. Thanks are also due to the Maine Office of Substance Abuse for their cooperation with this research. The responsibility for all errors and omissions remains with the authors.

1 Introduction

Relative performance evaluations provide incentives to workers and managers, identify best practices in organizations, and inform resource allocation decisions. In light of these merits, the neglected problem of implementing meaningful relative performance measurements is important. In this paper, we address the practical problem of comparing the performance of similar activities that are not directly comparable. Our specific application is the provision of alcohol-abuse treatment in the state of Maine (USA). Essentially, we develop a methodology for "benchmarking" the performance of these health service providers, and apply this methodology to estimate potential efficiency gains from identifying and transferring "best practices".

As a general motivation for our analysis, consider the following stylized problem. Suppose a decision-maker is interested in comparing the cost-effectiveness of N independent service units. The decision-maker observes costs and an output measure for each unit. However, the naive use of these data to compare productivities is problematic because the units produce differentiated services. Such difficulties of relative performance evaluation are particularly apparent for health care services. First, the activities of the service units may not be directly comparable because each is treating a different patient population. A unit may perform worse simply because of a more difficult case load. Second, the service units may be hard to compare because their activities differ in quality. For example, the outcome measure might be a measure of morbidity immediately following a hospital procedure. The discharged patients of a hospital with low measured morbidity might suffer future complications or have low survival rates. Under circumstances like these, how can the decision-maker meaningfully compare the performance of the N service units?

We confront the real life difficulties of comparing the performance of publicly-funded substance abuse treatment providers in the state of Maine in the United States. Maine's Office of Substance Abuse (OSA) in the 1980's and 90's allocated public funds to qualified substance-abuse treatment programs. OSA thought that some of these programs might be better than others at remedying drug and alcohol abuse, and collected data in order to measure and compare their performance. The inputs used in treatment and the costs of providing treatment were relatively easy to monitor, but the outputs of the treatment agencies were difficult to measure directly and difficult even to define precisely. Adding to this difficulty, the agencies treated populations living in different parts of the state. Although OSA measured characteristics of these different patient populations, it is likely that there is important unmeasured heterogeneity in these populations. We develop and apply a methodology for comparing the performance of the different agencies treating different populations. The methodology is based on a model of the treatment processes that is designed to exploit the information contained in multiple outcome measures. The outcome measures are time in treatment, completion of treatment, and alcohol use at completion.

To illustrate some of the issues that our methodology addresses, consider Figure 1. The figure shows the percentage of admitted patients that successfully completed treatment in each of the 12 months after admission for three treatment agencies. These completion densities are very different. For Agency A, the density peaks at 3 months, for B at 6 months, and for C at 8 months. Moreover, patients in A and C were noticeably more likely to complete treatment within twelve months than those at B. There are several possible explanations for A's apparently superior performance. First, A might provide superior treatment, leading to a higher probability of completion compared to B or quicker success compared

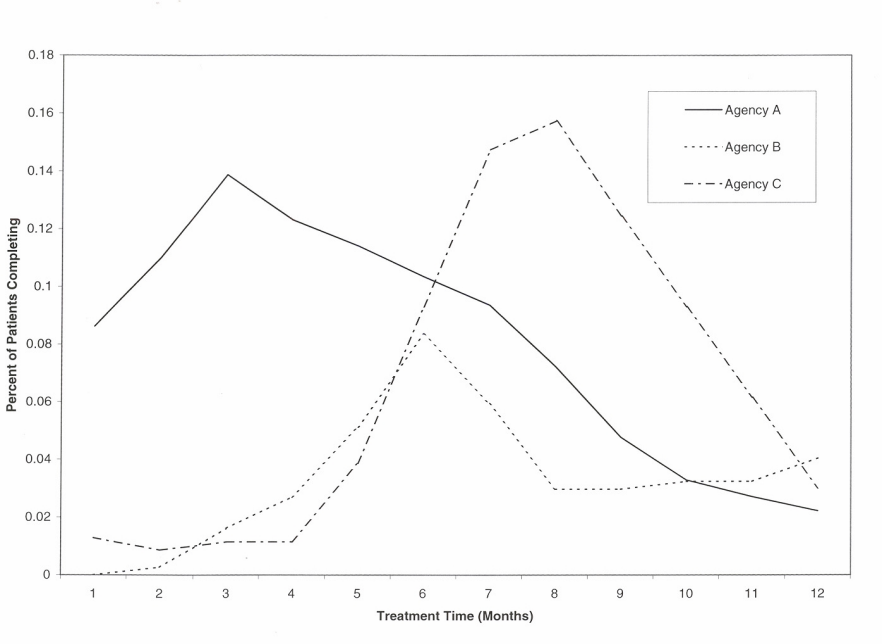

Figure 1: Sample Completion Densities

to C. Similarly, A might use more intense and therefore more costly treatment to achieve the better outcomes, e.g. provide patients more hours of treatment per week. Third, A's patients might enter treatment with less severe afflictions than patients at B and C. Lastly, A might have less stringent clinical standards for the completion of treatment. Obviously these different explanations have different implications for comparing the cost-effectiveness of the agencies.

To distinguish these alternative explanations empirically, we model a production process for a health treatment with the following general properties: (1) production takes time; (2) production is uncertain; (3) production occurs under varying exogenous conditions; (4) output is not perfectly observable. More specifically, we model the health treatment process as illustrated in Figure 2: a patient enters a treatment program with an initial (scalar-valued) health status (h_0), health status evolves stochastically through time in response to treatment

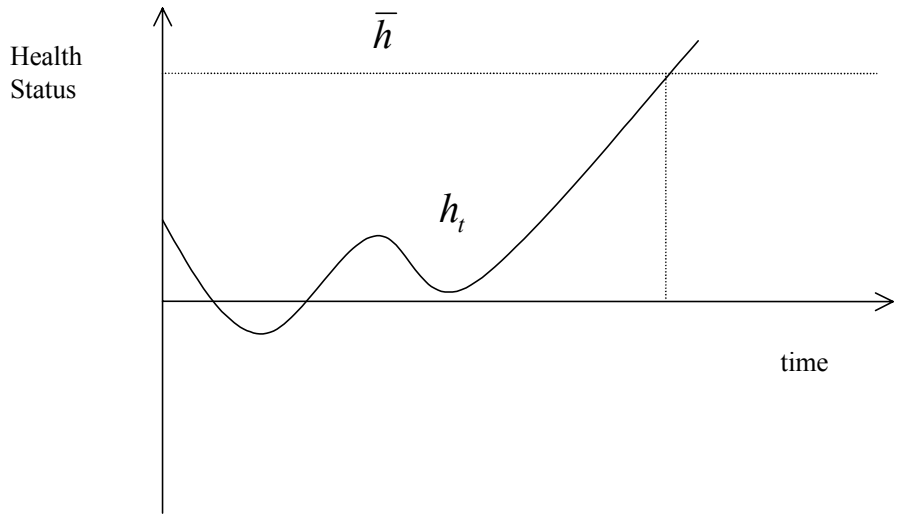

Figure 2: Health care treatment process

(h_t), and treatment is completed when the patient's health status crosses an upper threshold ($\overline{h}$). We model the evolution of h_t as a Wiener process, and completion of treatment as the first crossing of the barrier $\bar{h}$. As a result, the time it takes to complete treatment has an Inverse Gaussian distribution.

In order to deal with the measurement problems outlined above, our empirical implementation of the model allows that: (a) different treatment programs differ in their effectiveness at improving health status; (b) different programs have different populations of patients entering treatment (both with regard to observables and econometric unobservables); and (c) different programs have different thresholds for the completion of treatment. The empirical model also controls for the fact that some patients exit treatment prematurely, and allows the rate of attrition to vary from program to program. Our estimation methodology is maximum likelihood, using the Inverse Gaussian distribution for the time to completion

of treatment.[1] Health status is not directly observable, although the data set includes indicators of health status at admission and discharge. We are careful to control for possible unobserved heterogeneity and to distinguish "treatment effects" from unobserved population characteristics.

Our model is estimated with data drawn from an admission-discharge data set for patients receiving outpatient treatment for alcohol abuse provided by publicly-funded agencies in the state of Maine. This is the data that OSA collected to inform its budget allocation problem. The data set matches patients to treatment agencies. Our analysis indicate that these agencies differ substantially not only in their ability to improve the health of alcohol abusers, but also in their completion thresholds, the initial health of their patient populations, and their ability to retain patients until completion of treatment. The estimated model enables us to simulate the improvement in the health status of any population of patients in any treatment program. We also separately estimate an equation predicting the quantity of treatment patients would receive at different agencies, combine this with unit cost data drawn from the contracts between the treatment agencies and OSA, and thereby predict the total cost of treating any population of patients in any treatment program. Our results show significant differences in the estimated productivities, i.e. the improvement of health per dollar spent, of the different treatment agencies, suggesting the possibility of efficiency gains if the less productive agencies were to adopt the "technologies" (e.g. practice styles and management techniques) of the more efficient ones. In practice, such a technology transfer might be accomplished by mergers, information exchange, or training.

We estimate that a transfer of best practices could reduce treatment costs significantly

[1] See Lancaster (1990) for a discussion of the Wiener process as a duration model.

without compromising health outcomes. The nature of our policy experiments is illustrated in Table 1, which presents estimates (based on our estimated model) of how Maine could reduce expenditures while not reducing the number of abstinent discharged patients. For example, by transferring the practice style of Agency 6 to Agency 13's patient population, Maine could save a significant $63,260 (standard error in parentheses), and by transferring the practice style of Agency 12 to Agency 15's population, another $29,091. The total of such profitable transfers comes to $203,752, which is 9.27% of the budget of the top fifteen publicly funded agencies.

The recent literature on treatment effects has emphasized non-parametric or semi-parametric estimation to obtain robust conclusions (Manksi, 1996), and instrumental variables to control for endogenous selection (Imbens and Angrist (1994), Heckman (1997)). In our empirical context, good instruments to control for unobserved patient heterogeneity are not readily available. Therefore, we follow the structural literature (e.g. Olley and Pakes (1996), Keane and Wolpin (1997), and Eckstein and Wolpin (1999)) by specifying a model that identifies the parameters of interest. As detailed later, the key identification restrictions arise very naturally in our model of the treatment process. Our methodology is most related to that of Olley and Pakes (1996) who impose reasonable covariance restrictions to estimate a production function with endogenous inputs. In our case, an "endogeneity problem" arises because unobserved patient characteristics might be correlated with patients' assignments to treatment agencies.[2] In the same spirit as Olley and Pakes (1996), we exploit data on alcohol use at admission to help control for the resulting endogeneity problem. Finally, our parametric model has the benefit of enabling strong predictions from policy simulations, in particular the "transfer of best practices" policy experiment mentioned above.

[2] We also find direct evidence that patient populations differ in their unobserved characteristics.

Controlling for Unobservable Population Effects				
Population of Agency	Treated by Agency	$ Savings	Cumulative $ Savings	Cumulative % Savings
13	6	63259.59	63259.59 (11115.76)	0.0288 (0.0049)
15	12	29091.40	92350.99 (14170.04)	0.0420 (0.0063)
7	6	27515.10	119866.09 (21630.42)	0.0546 (0.0097)
10	4	26500.25	146366.34 (27846.17)	0.0666 (0.0126)
3	4	19480.57	165846.91 (44810.56)	0.0755 (0.0205)
14	12	16635.7	182482.61 (45690.58)	0.0831 (0.0209)
8	4	8958.95	191441.56 (80298.50)	0.0871 (0.0367)
11	12	6724.75	198166.31 (81374.71)	0.0902 (0.0372)
9	4	3368.79	201535.10 (83680.47)	0.0917 (0.0382)
1	2	2217.06	203752.16 (82687.48)	0.0927 (0.0378)
6	6	0	203752.16 (82687.48)	0.0927 (0.0378)
5	5	0	203752.16 (82687.48)	0.0927 (0.0378)
4	4	0	203752.16 (82687.48)	0.0927 (0.0378)
12	12	0	203752.16 (82687.48)	0.0927 (0.0378)
2	2	0	203752.16 (82687.48)	0.0927 (0.0378)

Table 1: Cost reducing technology transfers, controlling for population effects in terms of abstinence

The rest of the paper is organized as follows. Section 2 describes our basic model of the treatment process, and outlines our estimation methodology. Section 3 describes the data set to which we apply the model; Section 4 discusses the estimated model, applies the model to compare the performance of different treatment agencies and estimates the potential gains from transferring "best practices." Section 5 concludes and outlines some possible directions for further research. The appendix discusses additional estimation and simulation results and the health services literature on the effectiveness of substance abuse treatment.

2 Model and estimation strategy

2.1 Time in treatment and health improvement

Consider the following model of health care treatment. A patient i enters treatment with an initial health status h_{i0}. Once in treatment, the patient's health status, h_{it}, evolves stochastically according to a Wiener process with drift μ_i and variance σ^2. If h_{it} crosses an upper threshold, $\overline{h}_i$, then the patient is deemed to have completed treatment successfully. It follows from the distribution theory for Wiener processes that the cumulative probability distribution of completion times for patient i is the Inverse Gaussian distribution,

$$F(t_i; \mu_i, \sigma, \Delta h_i) = 1 - \Phi(\frac{\Delta h_i - \mu_i t}{\sigma\sqrt{t}}) + e^{\frac{2\mu_i \Delta h_i}{\sigma^2}} \Phi(\frac{-\Delta h_i - \mu_i t}{\sigma\sqrt{t}}) \tag{1}$$

where $\Delta h_i = \overline{h}_i - h_{i0}$ is the total change in health necessary for completion, and $\Phi(\bullet)$ is the cumulative normal distribution. A simple estimation strategy would be to fit time until completion of treatment (t_i) to the Inverse Gaussian distribution function by estimating μ_i, σ, and Δh_i as parametric functions of patient characteristics.

A problem with this simple approach is that some patients drop out of treatment before

completion. This is problematic because patient attrition is potentially a source of selection bias. For example, if less healthy patients are more likely to drop out before completion of treatment, then those agencies with a "slower" treatment process will have a healthier population of patients who complete treatment. One way to control for patient selection is as follows.[3] If $\lambda_i(t)$ is patient i's time-varying exogenous dropout hazard, then the distribution of dropout times for patient i is:

$$G(t_i; \lambda_i(\cdot)) = 1 - e^{-\int_0^{t_i} \lambda_i(s)ds}. \tag{2}$$

Therefore, if c_i is a dummy variable indicating whether the patient has completed treatment, rather than exiting prematurely, and t_i is the patient's time in treatment, then the likelihood function for (c_i, t_i) is

$$\{f(t_i; \mu_i, \sigma, \Delta h_i)[1 - G(t_i; \lambda_i(\cdot))]\}^{c_i} \times \{g(t_i; \lambda_i(\cdot))[1 - F(t_i; \mu_i, \sigma, \Delta h_i]\}^{1-c_i} \tag{3}$$

where $f(\cdot)$ and $g(\cdot)$ are the density functions derived from $F(\cdot)$ and $G(\cdot)$, respectively. An empirical implementation of this model might specify μ_i, σ, λ_i, and Δh_i as parametric functions of patient characteristics and estimate the parameters by maximizing the likelihood function. This strategy controls for patient attrition by exploiting two outcome measures: completion of treatment (c_i) and time in treatment.(t_i).[4]

A limitation of this more sophisticated model is that it does not separately identify the two components of $\Delta h_i = \overline{h}_i - h_{i0}$, and therefore cannot distinguish whether a treatment program has a higher completion threshold $(\overline{h}_i)$ or a healthier population at admission (h_{i0}).

[3] For different views on how to deal with dropouts in the substance abuse literature see, for example, Anglin and Hser (1990), Apsler (1991), Apsler and Harding (1991), Ball and Ross (1991), and the Gerstein and Hardwood (1990).

[4] A possibility for further research is to endogenize a patient's exit decision. For example, a patient's decision to leave treatment before completion might be triggered by a disappointing evolution of health status while in treatment.

Separate identification of these differences is possible if the model is augmented to incorporate additional outcome measures. Assume that there are measures of health observed at admission to treatment and at completion of treatment: q_{i0} is observed at admission and $\Omega_{i0}(q_{i0}|h_{i0})$ is the probability of q_{i0} conditional on h_{i0}; similarly, $\overline{q}_i$ is observed if the patient completes treatment ($c_i = 1$),[5] and $\overline{\Omega}_i(\overline{q}_i|q_{i0}, \overline{h}_i, h_{i0})$ is the conditional probability of $\overline{q}_i$ given $(q_{i0}, \overline{h}_i, h_{i0})$. Then the likelihood function for $(c_i, t_i, q_{i0}, \overline{q}_i)$ is

$$\{f(t_i; \mu_i, \sigma, \Delta h_i)[1 - G(t_i; \lambda_i(\cdot))]\}^{c_i} \times \{g(t; \lambda_i(\cdot))[1 - F(t_i; \mu_i, \sigma, \Delta h_i]\}^{1-c_i} \times$$
$$\Omega_{i0}(q_{i0}|h_{i0}) \times \bar{\Omega}_i(\overline{q}_i|q_{i0}; \overline{h}_i, h_{i0})^{c_i} \tag{4}$$

Taking logs and summing over patients gives the log likelihood function for the sample of patients under the assumption of independent likelihood functions across patients.

Our estimation strategy for the augmented model is to specify h_{i0}, $\overline{h}_i$, μ_i, $\lambda_i(t)$, $\Omega_{i0}(q_{i0}|h_{i0})$, and $\bar{\Omega}_i(\overline{q}_i|q_{i0}, \overline{h}_i, h_{i0})$ as functions of the data, substitute these functions into the sample log likelihood function, and maximize the resulting function to obtain parameter estimates. Importantly, we allow for an unobserved patient characteristic θ_i. Let a patient's initial health be a linear function of patient characteristics observed by the econometrician (vector $\mathbf{X}_i$) and an unobserved characteristic (θ_i), i.e.

$$h_{i0} = \alpha_0 \mathbf{X}_i + \theta_i \tag{5}$$

Let A_{ij} denote a dummy variable that is equal to 1 if patient i is assigned to program j $(= 1, ...J)$, and equal to 0 otherwise, and $\mathbf{A}_i$ the vector of these dummy variables for patient i. We specify the drift, threshold, and log dropout hazard as linear functions of these

[5] As is often the case, we have less reliable information on the status of patients who drop out of treatment than for those who complete treatment.

variables:

$$\mu_i = \delta_0 \mathbf{X}_i + \delta_1 \mathbf{A}_i + \delta_2 \theta_i \tag{6}$$

$$\overline{h}_i = \beta_0 \mathbf{X}_i + \beta_1 \mathbf{A}_i + \beta_2 \theta_i \tag{7}$$

$$\ln \lambda_i(t) = \gamma_0 \mathbf{X}_i + \gamma_1 \mathbf{A}_i + \gamma_2 \theta_i + \gamma_3 t \tag{8}$$

The agency dummies $\mathbf{A}_i$ enter these three equations directly. We refer to the coefficients on these dummies as "treatment effects," and interpret them as capturing differences in practice style across agencies that result in a different treatment process. This is a quasi-reduced form approach to estimating treatment effects in the sense that we are not modelling the underlying decisions that determine these effects. For example, a particular agency might have a higher than normal drift rate because the agency's practice style is to treat patients more intensely (e.g. schedule more hours of treatment per week). Perhaps the same agency also has higher attrition rates due to the more demanding schedule.[6] Importantly, $\mathbf{A}_i$ does not directly enter the h_{i0} equation. The reason is clear: h_{i0} is a variable determined prior to the treatment process and by definition not subject to treatment effects of the agencies.

If patients were assigned randomly to agencies, then we could treat θ_i as an unobservable independent of $\mathbf{A}_i$ and estimate the treatment effects $(\delta_1, \beta_1, \gamma_1)$. However, since this is not an experimental setting, it seems presumptuous to assume that θ_i is uncorrelated with $\mathbf{A}_i$.[7]

[6] Later, we relate agency performance to a few characteristics of the agency. An alternative approach would be to incorporate key agency characteristics directly into the model in lieu of the treatment effect dummies.

[7] In our alcohol abuse treatment setting, this correlation could arise because agency assignment is primarily based on where patients live and patients characteristics might vary geographically. In fact, observed characteristics $\mathbf{X}_i$ are clearly correlated with $\mathbf{A}_i$ in our data. As such, it is likely that the unobserved θ_i is correlated with $\mathbf{A}_i$.

There are also other possible reasons for correlation between $\mathbf{A}_i$ and θ_i. For example, it is conceivable that some treatment agencies selectively admit patients based on h_{i0}. Our model may not perfectly capture selection where patients are accepted or rejected based on h_{i0}, since in this case the distribution of θ_i conditional on $\mathbf{A}_i$ and $\mathbf{X}_i$ might be truncated. However, we should pick up the mean effect of such selection in our α_1.

Accordingly, we assume that the conditional distribution of θ_i given $\mathbf{A}_i$ and $\mathbf{X}_i$ is normal with mean $\pi_1\mathbf{X}_i + \alpha_1\mathbf{A}_i$, i.e.

$$\theta_i = \pi_1\mathbf{X}_i + \alpha_1\mathbf{A}_i + \varepsilon_i \tag{9}$$

where $\varepsilon_i \sim N(0,1)$.[8] Substituting (9) into (5) - (8) we obtain

$$h_{i0} = \alpha_0\mathbf{X}_i + \alpha_1\mathbf{A}_i + \varepsilon_i \tag{10}$$

$$\mu_i = \delta_0\mathbf{X}_i + (\delta_1 + \delta_2\alpha_1)\mathbf{A}_i + \delta_2\varepsilon_i \tag{11}$$

$$\overline{h}_i = \beta_0\mathbf{X}_i + (\beta_1 + \beta_2\alpha_1)\mathbf{A}_i + \beta_2\varepsilon_i \tag{12}$$

$$\ln\lambda_i(t) = \gamma_0\mathbf{X}_i + (\gamma_1 + \gamma_2\alpha_1)\mathbf{A}_i + \gamma_2\varepsilon_i + \gamma_3 t \tag{13}$$

where the coefficients on $\mathbf{X}_i$ have been redefined to include π_1.[9]

Examining (11)-(13), note that there are two effects of $\mathbf{A}_i$ on each of μ_i, $\overline{h}_i$ and λ_i. The first effect (e.g. the J-vector δ_1 in (11)) is the "treatment effect" - an attribute of the agency providing treatment to patient i. The second effect (e.g. the scalar δ_2 times the J-vector α_1) is the "population effect" - capturing the fact that a particular agency has patients with higher or lower values of the unobservable characteristic.

Separating these treatment effects and population effects is very important for our goal of comparing performances of different treatment agencies. We want to "rank" agencies based on the effectiveness of their treatment program (i.e. their treatment effects), and not

[8] Setting the variance of ε_i equal to unity is inoquous because health is a latent variable as it will become clear shortly. Hence, since h_{i0} is only "observed" through ordered variables representing frequency of alcohol use, we need to make this normalization to define what a "unit of health" is. Note that we are also currently assuming that the ε_i is homoscedastic, i.e. its variance does not depend on $(\mathbf{X}_i, \mathbf{A}_i)$.

[9] Precisely, the transformations are $\alpha_0 = \alpha_0 + \pi_1$, $\delta_0 = \delta_0 + \delta_2\pi_1$, $\beta_0 = \beta_0 + \beta_2\pi_1$, and $\gamma_0 = \gamma_0 + \gamma_2\pi_1$. This simply emphasizes that we cannot separate the direct effect of a particular $\mathbf{X}_i$ characteristic and the effect through that characteristic's correlation with the unobserved characteristic. This lack of identification muddies the interpretation of the coefficients on $\mathbf{X}_i$.

based on whether they happen to admit patients that are unobservably easier or harder to treat (i.e. their population effects). Importantly, these treatment and population effects are separately identified because, by definition, h_{i0} does not depend on a treatment effect. The observed effect of $\mathbf{A}_i$ on h_{i0} gives us a pure view on the population effect which can then be "netted" out of the other equations.

Moving to our indicators of health at admission and completion, let q_{i0} and $\overline{q}_i$ be categorical variables with $M+1$ possible integer values, indexed $m = 0, ...M$. We assume that q_{i0} and $\overline{q}_i$ are related to h_{i0} and $\overline{h}_i$ through an ordered probit structure. More specifically,

$$\begin{aligned} q_{i0} &= M \text{ iff } h_{i0} + \eta_{i0} < \psi_M \\ &= m \text{ iff } \psi_{m+1} \leq h_{i0} + \eta_{i0} < \psi_m, \text{ for } m = 1, ...M-1 \\ &= 0 \text{ otherwise} \end{aligned} \tag{14}$$

$$\begin{aligned} \overline{q}_i &= M \text{ iff } \overline{h}_i + \overline{\eta}_i < \psi_M \\ &= m \text{ iff } \psi_{m+1} \leq \overline{h}_i + \overline{\eta}_i < \psi_m, \text{ for } m = 1, ...M-1 \\ &= 0 \text{ otherwise} \end{aligned} \tag{15}$$

where

$$\begin{pmatrix} \eta_{i0} \\ \overline{\eta}_i \end{pmatrix} \sim N\left(0, \begin{bmatrix} \sigma_\eta^2 & \sigma_{0T} \\ \sigma_{0T} & \sigma_\eta^2 \end{bmatrix}\right) \tag{16}$$

i.e., the unobservables η_{i0} and $\overline{\eta}_i$ are distributed joint normal with common variance σ_η^2 and with covariance σ_{0T}.[10]

Conditional on h_{i0}, $\overline{h}_i$, μ_i, and $\lambda_i(t)$, the likelihood function for patient i is given by

[10] This bivariate normal distribution defines $\Omega_{i0}(q_{i0}|h_{i0})$, and $\bar{\Omega}_i(\overline{q}_i|q_{i0}, \overline{h}_i, h_{i0})$ in (4). In actual computation of the likelihood function it is easier to work directly with the joint distribution $\Omega_i(\overline{q}_i, q_{i0}|\overline{h}_i, h_{i0})$. The equality of the variance terms is a simplification.

the "analytically" computable[11] equation (4) above. However, h_{i0}, $\overline{h}_i$, μ_i, and $\lambda_i(t)$ are unobserved as they depend on the unobserved characteristic ε_i. To compute the likelihood, it is necessary to integrate (4) over the distribution of ε_i. As there is no analytic solution for this one-dimensional integral, we simulate it.[12] Maximization of the simulated log likelihood function returns estimates for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \boldsymbol{\psi}, \sigma, \sigma_\eta^2, \sigma_{0T})$ using data $(c_i, t_i, q_{i0}, \overline{q}_i, \mathbf{X}_i, \mathbf{A}_i)$ for a sample of patients $i = 1, ...n$.

Several clarifications and caveats about the model are in order. First, the four separate sets of coefficients on $\mathbf{X}_i$ and $\mathbf{A}_i$ ($\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$, and δ) are identified by the four endogenous variables. The coefficients in h_{i0} are pinned down by the observed indicators of initial health, q_{i0}. Likewise, the coefficients in $\overline{h}_i$ are identified from the observed indicators of final health $\overline{q}_i$. The coefficients in μ_i and $\lambda_i(t)$ are jointly identified by the completion (c_i) and time in treatment (t_i) data. Increases in $\lambda_i(t)$ decrease the probability of completion. Increases in μ_i increase the probability of completion and decrease the expected time until completion (conditional on completion).[13] Second, $\mathbf{X}_i$ and θ_i do not enter the ordered probit relationship between the h's and the q's. This is a necessary normalization because coefficients on $\mathbf{X}_i$ in the q equations would not be identified separately from those in the equations for h_{i0} and $\overline{h}_i$. The coefficients in h_{i0} and $\overline{h}_i$ must be interpreted accordingly.[14] Third, we suppose that

[11] $\Omega_i(\overline{q}_i, q_{i0}|\overline{h}_i, h_{i0})$ is computed numerically using highly accurate bivariate normal CDFs.

[12] We use an 11-point discretized normal approximation to the distribution of ε_i.

[13] We expect that even more elaborate models can be estimated satisfactorily. For example, consider allowing $\mathbf{X}_i$ and $\mathbf{A}_i$ to affect the variance of the Wiener process (σ^2). These effects are separately identified because a change in σ^2 has a different effect on the variance of time until completion than does a change in μ_i.

[14] As described in more detail later, q_{i0} and $\overline{q}_i$ are measures of frequency of alcohol consumption (e.g. once a month, once a week, etc.). Patient characteristics might enter the relationship between h and q if the relation between health and how much one drinks is different for different patients. Importantly, the potential inclusion of θ_i in the relationship between h_{i0} and q_{i0} does not hinder our separate identification of population effects and treatment effects. On the other hand, the restriction that A_i does not directly (it can indirectly enter through θ_i) enter this relationship is crucial for this separate identification of population and treatment effects. Without this restriction, this implicit normalization would contradict our interpretation of the agency effect in the equation for h_{i0} as a pure "population effect." We think this is a reasonable

the health indicator at discharge, $\overline{q}_i$, is observed only if the patient completes treatment.[15] Fourth, we measure time in treatment in weeks, and aggregate the likelihoods accordingly. Fifth, when we discuss the data in the next section, we will modify our model slightly to account for patients who are still in treatment at the end of the sample period and to better account for the empirical time pattern of attrition.

Beside providing a concise and realistic framework for combining information from multiple endogenous outcomes $(c_i, t_i, q_{i0}, \overline{q}_i)$, our structural model solves the endogeneity problem resulting from a potential correlation between the vector of agency dummies $\mathbf{A}_i$ and the unobserved characteristic ε_i. This solution relies on two key assumptions. First, as discussed above, the dependent variable q_{i0} is free of treatment effects. This allows us to estimate relative population effects from the q_{i0} equation. Second, unobserved patient characteristics are summarized by a scalar.[16] This allows us to assess the contribution of the population effects to each dependent variable by conditional covariances of the four variables.[17,18]

restriction because q is an objective measure of a patient's condition.

[15] We do have some data on health status at discharge for non-completors. We are concerned, however, about its reliability and are not using it in estimation.

[16] While this scalar restriction is significant, it should capture the first order effects of unobserved characteristics and is certainly preferrable to ignoring unobserved characteristics completely.

[17] Slightly more formally, we have four dependent variables $(c_i, t_i, q_{i0}, \overline{q}_i)$. There are five unobservables in the model: the Wiener process, the dropout process, the two error terms in the relation between h_{i0} $(\overline{h}_i)$ and q_{i0} $(\overline{q}_i)$, and the unobserved patient characteristics. These unobservables are assumed uncorrelated except for correlation between the two unobservables in the $q-h$ relationship. Thus, there are three variances (variances of the unobservable characteristic (ε) and the dropout process are normalized to 1), one covariance, and three parameters (δ_2, β_2 and γ_2) to be identified by conditional covariances of the four dependent variables. The eight free elements (two variances must be normalized due to discreteness) in this four by four conditional covariance matrix appear to be sufficient to identify these seven parameters. Thus, identification of the model (ignoring selection issues - see next footnote), comes from conditional means and variances rather than higher order distributional assumptions.

[18] While the argument in the previous footnote suggests that identification of our basic model comes from means and variances and doesn't rely on precise functional forms for our distributions, there are two data problems in our particular application that do bring distributional issues into the identification picture. First, as is standard in discrete choice models, the discreteness of q_{i0} and $\overline{q}_i$ requires distributional assumptions that do affect identification. An additional issue arises because we only observe $\overline{q}_i$ for completors. This generates a selection problem. Our model implicitly deals with this selection problem in a Heckman-like correction. As is usually the case in these selection models, it is likely that the correction *does* depend on

Importantly, this endogeneity issue in our data set would be very hard to deal with in less structural models. The standard approach would look for instruments that were correlated with agency assignment yet uncorrelated with unobserved patient characteristics.[19] It is hard to imagine such instruments existing if the endogeneity is due to inherent differences in patient populations across regions and treatment agencies have different geographical footprints.[20]

There is also a second possible "endogeneity" problem arising from the fact that "time in treatment" (t_i) and "health improvement" ($\triangle h_i$) are determined simultaneously. To illustrate the problem, consider the hypothetical question: "If a patient were required to stay in treatment an extra week, what would be the expected increase in the patient's health?"[21] One way to answer this question is to estimate a causal "production function", e.g.

$$\triangle h_i = \mu t_i + \beta \mathbf{X}_i + \varpi \mathbf{A}_i + \varsigma_i \tag{17}$$

where changes in health are caused by time in treatment. A primary problem in this approach is that the unobservables ς_i may be correlated with t_i. Intuitively, patients that are

the distributional assumptions in the model (e.g. the normality of the unobserved patient characteristics and the Wiener process).

[19] This discussion highlights two general approaches to deal with endogeneity. The first and most common approach is instrumental variables. The second approach, taken here and following Olley and Pakes (1996), is to use auxiliary equations/data to, in essence, "observe" the unobservables causing the endogeneity problem.

[20] More precisely, suppose, for example, that agency assignment is purely regional, e.g. one goes to the closest agency. Suppose also that there are unobservable differences in patient populations across regions (our empirical results suggest that this is true). Then there would be literally *no* possible instruments for the endogenous agency assignment. Moreover, even if there were other determinants of agency assigment they may not be feasible instruments. For example, agency budgets may be perceived as good candidates for instrumenting agency assigment. There are, however, at least two reasons to discard them. First, one might expect the budget to affect agency decisions on drift and thresholds, in which case they would not be sufficient to control for the endogeneity. Second, OSA may have information about the unobservable patient characteristics and incorporate them in their budget allocation decision, implying correlation between budget and ε.

[21] Or perhaps, "if we wanted to increase health to a given level, how long would the patient likely have to stay?"

more easily treated take less time to complete treatment due to unobservable characteristics that matter for the treatment process. A standard approach to this endogeneity problem is to identify excluded variables that serve as instruments for t_i.[22] In contrast to the usual instrumental variables approach based on exclusion restrictions, our structural model consistently estimates the health production function based on restrictions that arise naturally from the assumed treatment process. In particular, the assumption that health improvements accrue over time restricts how the variables affecting the drift of the Wiener process interact with time in treatment.[23]

2.2 Quantity of treatment

We predict the total cost of treatment for various patient populations in various agencies with a separate estimation procedure. Given data on agency costs per unit of treatment, a model of the number of units of treatment provided to a patient is sufficient to estimate the cost of treatment. We assume that the units of treatment[24] individual i receives, u_i, is given

[22] The unobservability of $\triangle h_i$ might be dealt with by replacing it with an observable proxy for health improvement (e.g. reduction in use, $\overline{q}_i - q_{i0}$) and estimate a probit or logit type model. Lu and McGuire (2001) employ an approach like this, using the same data set as we do. They use both "units of treatment" and "time in treatment" as explanatory variables and agency revenue sources as instruments to control for the endogeneity of these variables. They do not explicitly address the endogeneity of agency assignment and, therefore, do not separately identify population from treatment effects.

[23] According to our Wiener process model, the production function is:

$$\begin{aligned}\triangle h_i &= \mu_i t_i + v_i = (\delta_0 \mathbf{X}_i + \delta_1 \mathbf{A}_i + \delta_2 \theta_i) t_i + v_i \\ &= \delta_0 \mathbf{X}_i t_i + (\delta_1 + \delta_2 \alpha_1) \mathbf{A}_i t_i + \delta_2 \varepsilon_i t_i + v_i.\end{aligned}$$

This means that if we could observe $\triangle h_i$ then we would get consistent estimates by dividing through by t and using OLS, or, alternatively, by using $\mathbf{X}$ and $\mathbf{A}$ as instruments for $\mathbf{X}t$ and $\mathbf{A}t$, respectively. Note, however, that this simple approach does not separately identify population from treatment effects.

[24] In expressing the total units a patient receives in individual therapy units we assume that one unit of group therapy is equivalent to w_g units of individual therapy where w_g is the ratio between group therapy unit costs and individual treatment unit costs. Similarly, one unit of family therapy is equivalent to w_f units of individual therapy where w_f is the ratio between family therapy unit costs and individual treatment unit costs. The left hand side variable is therefore decomposed as follows, where lower case are the actual number

by:

$$\ln u_i = \xi_0 + \xi_1 \mathbf{X}_i + \xi_2 \mathbf{A}_i + \xi_3 \theta_i + \varkappa_i \tag{18}$$

$$= \xi_0 + \xi_1 \mathbf{X}_i + (\xi_2 + \xi_3 \alpha_1) \mathbf{A}_i + \xi_3 \varepsilon_i + \varkappa_i \tag{19}$$

where[25] $\varkappa_i$ is an unobservable term and ξ_2 is a vector of agency treatment effects. $\varkappa_i$ is naturally correlated with the unobservables governing the treatment process and/or the dropout process. For example, patients that drop out prematurely will also tend to receive less treatment. We can easily obtain the coefficients ξ_1 and $(\xi_2 + \xi_3 \alpha_1)$ with an OLS regression, but this does not separately identify the treatment effect ξ_2 and the population effect $\xi_3 \alpha_1$. This is unsatisfactory for our purposes. As in the other parts of our model, when we perform the conceptual experiment of moving patients into a different treatment program, the population effect should be held constant while the treatment effect should change.

We adopt the following approach to identifying the treatment and population effects for the units equation. From the estimated treatment model we recover a posterior mean for the unobservable characteristic ε_i for each patient i in the sample. This posterior is computed conditional on the estimated parameters $\widehat{\beta}$, the exogenous data X_i and A_i, and the observed frequency of use at admission q_{i0}. Call this posterior $\widehat{p}_i$. Since by definition

$$E[\varepsilon_i | X_i, A_i, q_{i0}, \widehat{\beta}] = \widehat{p}_i \tag{20}$$

it follows that:

$$\varepsilon_i = \widehat{p}_i + \varphi_i \tag{21}$$

of patient units received by patient i as reported in the MATS discharge form:

$$\ln u_i = \ln(u_{ii} + w_g \times u_{ig} + w_f \times u_{if}).$$

[25] Again we have redefined the coefficients on $\mathbf{X}_i$ to include potential correlation between observed and unobserved characteristics.

where φ_i is mean independent of X_i, A_i, and $\widehat{p}_i$. Substituting this expression for ε_i into the units equations yields the estimating equation:

$$\ln u_i = \xi_0 + \xi_1 \mathbf{X}_i + (\xi_2 + \xi_3 \alpha_1)\mathbf{A}_i + \xi_3 \widehat{p}_i + \xi_3 \varphi_i + \varkappa_i \tag{22}$$

By treating $\widehat{p}_i$ as data, it is possible to estimate $(\xi_2 + \xi_3\alpha_1)$ and ξ_3 by OLS, and then use the estimates of α_1 from section 4 to calculate the treatment effects ξ_2 in the units equation.

In this procedure, it is important to compute the posterior mean for ε_i by conditioning *only* on the initial condition q_{i0} and not on the endogenous variables $(c_i, t_i, \overline{q}_i)$. The reason is that $\varkappa_i$ likely is correlated with these endogenous variables because of correlation with the unobservables in the treatment and dropout processes (e.g. random missed appointments both decrease units of treatment and slow health improvement). There is not a similar problem with conditioning on q_{i0} because this variable is determined prior to the treatment process. Thus, when we only condition on q_{i0}, it is reasonable to assume that p_i is uncorrelated with $\varkappa_i$.

An issue that arises when we estimate the units equation separately from the production function is that we do not directly estimate covariances between estimated parameters of the production process and the parameters in the units equation. These covariances are necessary for obtaining standard errors on productivities and the results of our policy analyses that follow. To obtain these covariances, we bootstrap the estimation routine on resampled data sets and compute the covariances of the parameters across the bootstraps.[26]

[26] This procedure is less burdensome computationally than the alternative procedure of estimating the units equation simultaneously with the rest of the model.

3 The data

We apply our framework to treatment for alcohol abuse. Our primary data source is the Maine Addiction and Treatment System (MATS). The data describes people receiving outpatient treatment for alcohol abuse in the State of Maine (U.S.A.) between October 1990 and June 1996. Maine is mostly rural – in 1994, 35.9% of the population lived in metropolitan areas compared to 79.8% nationally.[27] Parts of the state are particularly remote. Substance abuse, and especially alcohol abuse, appears to be a serious problem in the state. Almost one percent of the population receives treatment in state funded substance abuse programs,[28] and for 87% of the adults who receive treatment, alcohol is the primary abused substance.[29] State government funding for substance abuse treatment was in the neighborhood of $10 million in 1995,[30] or about $8 per capita.[31]

MATS collects data on patients receiving treatment from any publicly funded substance abuse treatment agency in the state. MATS agencies were required to complete standardized admission and discharge forms for every treated patient, and to report this data to the state regulatory agency (OSA).[32] Our analysis focuses on the fifteen largest publicly funded substance abuse agencies, who had a combined budget of $\$6,439,312$ in fiscal year 1994 of which $\$4,594,595$ was budgeted for outpatient treatment alone.[33] Our treatment provider

[27] Data are from the census website: http://www.census.gov/statab/www/states/me.txt.

[28] Source: Maine's application for the Alcohol, Drug Abuse and Mental Health block grants in FY 1992

[29] Source: Maine OSA Fact Sheet, "Alcohol Use in Maine," 1997.

[30] The total budegeted expenditures of the Office of Substance Abuse was $10,085,716. Source: State of Maine Budget Document, 1994-95.

[31] Maine's population in 1990 was 1.228 million.

[32] The Department of Human Services was the relevant agency prior to the creation of OSA. In July 1990, OSA was created as a branch of the State's Executive Department. After July 1, 1996, OSA was transferred to the Department of Mental Health, Mental Retardation, and Substance Abuse Services (see: http://www.state.me.us/sos/cec/rcn/apa/depts.htm.) OSA is responsible for allocating state and federal funds for substance abuse, and for contracting with the agencies receiving these funds.

[33] We ignore the agencies that exclusively provide treatment under the Driver Education and Evaluation

variable A_{ij} is equal to one if patient i is treated at that agency ($j = 1, ...15$).

Substance abuse patients differ with respect to diagnosis and treatment modality. Our observation unit is a treatment episode, defined by an admission into treatment.[34] We focus on outpatients whose primary diagnosis is alcohol abuse,[35] and further restrict the sample to admissions in the four year period beginning October 1991. We also exclude repeat episodes, focusing on first time clients.

The resulting sample consists of 7601 treatment episodes. Of these, 3402 patients completed treatment, 4040 left treatment prematurely for one reason or another, and 159 were still in treatment at the end of the sample. To include the on-going patients in our econometric model, we modify the likelihood function in (4) to:

$$\{\{f(t_i;\mu_i,\sigma,\Delta h_i)[1-G(t_i;\lambda_i(\cdot))]\}^{c_i} \times \{g(t;\lambda_i(\cdot))[1-F(t_i;\mu_i,\sigma,\Delta h_i]\}^{1-c_i}$$
$$\times\Omega_{i0}(q_{i0}|h_{i0}) \times \bar{\Omega}_i(\bar{q}_i|q_{i0};\bar{h}_i,h_{i0})^{c_i}\}^{d_i} \times \{[1-G(t_i;\lambda_i(\cdot))][1-F(t_i;\mu_i,\sigma,\Delta h_i]\}^{1-d_i} \quad (23)$$

where $d_i = 1$ if the patient has been discharged from treatment, and $d_i = 0$ if the patient is still in treatment by the end of the sample. Notice that a patient who completes treatment ($c_i = 1$) or leaves treatment prematurely ($c_i = 0$) is automatically discharged ($d_i = 1$). Table

Program (DEEP). While these agencies treat large numbers of patients, they are intrinsically different from regular outpatient treatment. For example, DEEP patients need to complete a certain number of sessions to recover their licenses. This would entail a very different model than the one we describe in Section 2.

[34] We have also defined as a unique episode all outpatient episodes pertaining to the same patient that were concurrent in time or less than one month apart. In any of those cases we added the units of treatment received under each episode and used as admission information the admission information of the first episode and as discharge information the discharge information of the last episode. In most cases the different episodes belonged to the same agency but in those cases where the agencies were different we attributed the merged episode to the agency where the patient stayed longer.

[35] Outpatient treatment is essentially therapy. From the MATS data we distinguish individual, group, and family therapy as separate modalities. While the MATS data also distinguishes "intensive outpatient" and "adolescent outpatient" as distinct modalities, we include these in the category of individual therapy.

The MATS forms distinguish the treatment of primary clients from the treatment of codependents. A codependent is someone who receives treatment to better deal with someone else's substance abuse problem. We restrict attention to primary alcohol abusers receiving outpatient treatment. A primary patient is someone receiving treatment for their own substance abuse problem.

Average	Min	Max	Std
506.7	101	1388	422.8

Table 2: Statistics on number of patients per agency

Frequency of use at:		completion		
		abstinent	$<$ once daily	$\geq$ once daily
	abstinent (obs=1967)	0.990	0.010	0.000
admission	$<$ once daily (obs=1277)	0.856	0.144	0.000
	$\geq$ once daily (obs=158)	0.968	0.006	0.025

Table 3: Sample transition matrix for completors

1 presents some statistics on number of episodes handled by the fifteen different treatment agencies.

Three outcome variables are constructed from the MATS data. First, the discharge form distinguishes patients who completed treatment from patients who left treatment for other reasons.[36] From this, we define a dummy variable (c_i) that is equal to one if the patient completes treatment. Second, from the dates of admission and discharge we construct each patient's time in treatment (t_i). Third, the MATS forms report the frequency of alcohol use at both the time of admission and discharge. We define two categorical variables ($q_{i0}, \overline{q}_i$) to measure frequency of use at admission and completion respectively. These variables equal zero if the patient has been abstinent for at least one month, one if he currently uses alcohol less than once a day, and two if his frequency of use is even greater. The sample transition matrix for frequency of alcohol use is presented in Table 3.

Table 4 presents some summary data on completion of treatment, time in treatment, and frequency of use at both admission and discharge, in the aggregate population and for the

[36] Other reasons for leaving treatment are: 1) client leaves treatment without explanation or refuses treatment; 2) client cannot come to treatment either because of imprisonment or death; 3) and lastly the client is discharged from treatment due to non-compliance with rules and regulations.

			Population		Agency			
obs=7601			mean	s.d.	mean	min	max	s.d.
Status at discharge	% completion ($c = 1$)		44.8	49.7	44.0	31.7	59.4	9.5
	% premature exit ($c = 0$)		53.2	49.9	52.3	22.7	65.6	12.2
	% did not finish		2.1	14.3	3.7	0.5	18.0	4.9
Time in treatment	all		14.7	16.9	15.7	11.7	23.1	3.5
	$c = 1$		19.0	18.2	20.7	13.3	38.0	6.7
	$c = 0$		11.1	14.7	12.3	8.2	20.7	3.2
Frequency of use at completion	$c = 1$	abstinent	93.9	24.0	94.8	84.1	99.2	4.4
		$<$ once daily	6.0	23.8	5.1	0.8	15.9	4.4
		$\geq$ once daily	0.1	3.4	0.1	0.0	1.0	0.3
Frequency of use at admission	all	abstinent	50.2	50.0	48.3	34.7	67.8	10.7
		$<$ once daily	39.1	48.8	40.0	25.5	56.3	8.9
		$\geq$ once daily	10.7	30.9	11.7	6.2	20.8	4.8
	$c = 1$	abstinent	57.8	49.4	55.8	38.2	70.7	9.6
		$<$ once daily	37.5	48.4	38.8	25.5	56.6	8.8
		$\geq$ once daily	4.6	21.1	5.3	1.8	11.3	3.2
	$c = 0$	abstinent	43.9	49.6	42.1	25.0	65.8	11.9
		$<$ once daily	40.4	49.1	41.4	21.1	58.6	10.6
		$\geq$ once daily	15.7	36.4	16.4	10.3	23.3	4.3
Units of treatment	all		8.6	10.4	9.5	6.2	14.9	2.5
	$c = 1$		11.0	11.4	12.5	8.4	22.7	4.3
	$c = 0$		6.6	9.0	7.6	4.8	11.7	2.1

Table 4: Dependent variables statistics

fifteen agencies. 44.8% of the patients completed treatment successfully ($c = 1$), and 53.2% left treatment prematurely while only 2.1% of the patients are still in treatment by the end of the sample period. However, the patient populations at different agencies had very different average experiences. At the worst agency 31.7% of patients completed treatment, while at best 59.4% completed treatment. The average agency had 44% completion rate, and the standard deviation across agencies was 9.5%.

Similar agency heterogeneity is apparent for the average time patients spent in treatment and for the probability that a patient is abstinent at the time of discharge. The average

patient in the sample spent just under 15 weeks in treatment, but this average experience could vary between just under 12 weeks and just over 23 weeks depending on the agency. 93.9% of patients completing treatment were abstinent at the time of discharge but this success indicator could be as low as 84.1% or as high as 99.2% depending on the agency. Our econometric methodology is intended to exploit this heterogeneity in outcomes across agencies.

It is striking that a high percentage of patients enter treatment reporting that they have been abstinent over the past month. While a significant number of the patients come directly out of the penal system (plausibly explaining their abstinence), it is certainly conceivable that some patients may be misrepresenting their alcohol consumption prior to admission. For such "measurement error" not to undermine interpretations of empirical results, we must assume the underreporting does not vary across clinics conditional on observed patient characteristics $\mathbf{X}$.[37]

The bar chart in Figure 3 shows attrition for the entire sample population. These are the number of patients in each week of treatment that left prematurely. Clearly, there is an extraordinarily high attrition rate in the first week. After that, the empirical hazard is approximately constant. To better account for this pattern we modify our attrition model to allow for a higher probability of dropping out in the first week:

$$\ln \lambda_{i1} = [\gamma_0 \mathbf{X}_i + \gamma_1 \mathbf{A}_i + \gamma_2 \theta_i + \gamma_3]\gamma_5 + \gamma_4 \tag{24}$$

The extra coefficients γ_4 and γ_5 allow the hazard rate to jump discretely after the first week. Other aspects of our model are the same as described earlier.

[37] In other words, the underreporting of frequency of use at admission can vary across clinics, but this either be constant across clinics or be captured by the observables X. Frequency of use at admission is recorded on the admission form, and in the later part of the sample again on the discharge form. We use the former. By an oversight the latter was not extracted from MATS, which prevents us from checking reliability.

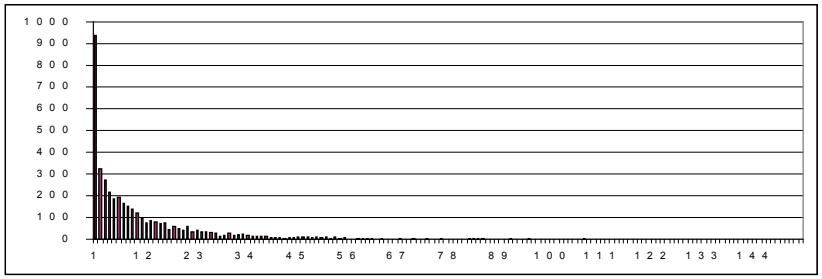

Figure 3: Sample attrition

MATS also reports a number of patient characteristics that we include in X_i. Table 5 exhibits summary statistics of these characteristics. 40.2% of patients are male, although some agencies have predominantly male populations and others the opposite. The average patient is thirty-something with a monthly household income of about \$871[38] a month. Average age is similar across the different agency populations, but there are clear differences in the average incomes.

We also include the source of payment for treatment as a patient characteristic. This could be important for several reasons. The fact that treatment is paid out of pocket, for example, could influence a patient's decision to exit treatment prematurely. Source of payment also impacts reimbursement amounts for agencies, and thus could influence an agency's determination of the completion time and/or the intensity of treatment. 35.1 % of patients in the sample pay for treatment out of pocket, 17.0% have private insurance, and another 23.9% are covered by Medicaid or Medicare. Treatment costs for 18.9% of patients in the sample nominally are covered by public funds allocated to the agency by OSA. Treatment costs for the small remainder of patients are paid by various other forms of public assistance and insurance. There is significant heterogeneity of revenue sources across agencies. It is

[38]This is the average monthly income when income is known and stated by the patient and is not bigger than \$9999.

also noteworthy that the majority of patients are involved with the legal system in some way. 21.3% are on parole or probation, and 28.5% are in treatment in connection with drunk-driving violations.

The last set of patient characteristics are variables relating more directly to a patient's health status at admission. First, we construct dummy variables from the clinician's assessment of the severity of the patient's condition, indicating whether the patient is a "casual," "life-style involved," "life-style dependent," or "dysfunctional" user, or if the patient's severity is "undetermined." Second, we construct dummy variables indicating whether in the past month the patient's drinking caused problems on the job or at school "not at all," "infrequently," "occasionally," or "frequently." Third, we construct a dummy variable indicating whether the patient has a concurrent psychiatric condition. All of these dummy variables are included in our list of patient characteristics ($\mathbf{X}$). The interpretation here is that these variables are direct predictors of unobserved health status h_{i0}. In particular, if our observed severity variable was continuous rather than reported in a coarse ordering, we might expect it to perfectly measure h_{i0} (or perhaps a linear combination of severity, problems on the jobs and psychiatric condition might perfectly measure h_{i0}). We also allow the more exogenous characteristics, i.e. the socioeconomic characteristics, to enter h_{i0} to the extent that the discretized variables do not perfectly measure unobserved health. This is important to remember in interpreting the coefficients on the socioeconomic characteristics - the coefficient on income, for example, measures the impact of income *conditional* on our observed severity measure.

We obtain unit cost data from a separate data set on each agency's contracts with OSA. Each agency's annual contract states budgeted total cost and budgeted units of treatment

obs=7601		Population		Agency			
		Mean	Std.	Mean	Min	Max	Std
Male		40.2	49.0	35.6	1.6	91.1	28.8
Age		33.1	10.8	33.1	30.3	36.2	1.5
Married		23.6	42.4	24.0	14.8	28.3	3.9
Income ($\leq$\$9999)		870.9	885.1	906.8	631.4	1172.3	170.7
Unemployed		32.2	46.7	30.3	17.4	40.4	7.2
Payer	OSA	18.9	39.2	17.4	0.8	54.3	15.7
	self	35.1	47.7	34.5	8.6	61.2	15.3
	private ins.	17.0	37.6	16.4	4.7	25.0	6.3
	medicaid/care	23.9	42.6	24.5	15.6	39.5	6.1
	other	5.1	21.9	7.2	1.7	18.8	4.5
Legal Inv.	none	41.3	49.2	42.6	18.2	75.3	13.9
	prob./parole	21.3	40.9	21.8	9.0	45.1	10.6
	drunk driving	28.5	45.1	26.4	1.0	49.2	12.6
	other	8.9	28.5	9.2	3.8	16.1	3.4
Adm. date	days	11923	394	11955	11620	12193	152
Quarter	I	20.2	40.2	20.4	10.9	30.7	5.4
	II	25.4	43.5	25.8	20.3	31.0	3.0
	III	27.6	44.7	27.1	17.8	32.8	4.2
	IV	26.7	44.3	26.8	20.5	36.7	3.6
Severity	casual/exp.	23.4	42.3	23.8	13.2	44.5	10.0
	involved	17.3	37.8	16.6	4.5	25.8	5.7
	dependent	40.5	49.5	42.2	29.5	63.4	7.6
	dysfunctional	18.4	38.8	16.8	0.0	43.3	12.3
	undetermined	0.4	6.4	0.5	0.0	2.1	0.7
Job problems	none	88.5	31.9	86.0	60.4	95.8	96.1
	infrequent	6.3	24.3	7.3	1.0	14.9	4.3
	occasional	2.7	16.3	3.9	1.2	13.9	3.5
	frequent	2.4	15.4	2.8	0.6	10.9	2.6
Pyschiatric problems		11.7	32.1	14.4	5.8	41.1	10.0

Table 5: Patient characteristics

for each treatment modality (e.g. group treatment is typically much less expensive than individual treatment). This gives us unit costs for individual, group and family therapy for each agency. We use relative unit costs to weight units of group and family therapy appropriately and construct the variable "total adjusted units of treatment" for each patient. This serves as the dependent variable in the units of treatment equation.[39]

Total costs for the policy simulations in subsection 4.3 are then constructed by multiplying predicted "adjusted units of treatment" by the unit cost of individual therapy. In these simulations we hypothesize that all patients were treated in 1993. 1993 unit costs for each agency were calculated by deflating the nominal unit cost of individual therapy by a 1992 county hospital wage index.[40]

4 Estimation results

4.1 Population and treatment effects

We focus primarily on our estimates of the treatment effects in the different agencies. These measure how the production of health varies between agencies *after* controlling for both observed and unobserved characteristics of agency populations. Our estimated coefficients on the X's are discussed in the appendix.

Table 6 presents estimated coefficients and standard errors for the various population and

[39] The calculated weights for group treatment were based on 1990-91 contract data, and for family treatment 1990-94 contract data. It was necessary to "clean" the data with various approximations to deal with the problems of missing observations, breaks in the series, and other problems.

[40] The cost deflator is actually a patient-weighted average hospital wage index that accounts appropriately for multicounty operations of some agencies. We lacked hospital wages for Sagadahoc county, and estimated the deflator for this county by the population-weighted deflator of the four neighboring counties.

treatment effects.[41,42] As discussed earlier, the coefficients in the h_0 equation are "population effects," measuring how patient populations of the agencies differ on average in their unobserved characteristics. The coefficients reported for μ, $\bar{h}$, and λ are the "treatment effects," e.g. for the drift equation, we report δ_1 rather than the total agency effect $(\delta_1 + \delta_2\alpha_1)$. The reported estimates are expressed as deviations from the population weighted mean, enabling a clear interpretation of statistically significant coefficients. For example, the fact that the treatment effect in the drift equation of Agency 1 is positive and significant indicates that a patient who remains in treatment at Agency 1 achieves a given completion threshold quicker than at a randomly assigned agency with the probability of assignment to an agency equal to the agency's population share.

First, the significance of nine of the fifteen population effects in the first column suggests that these agencies *do* differ in the unobserved characteristics of their populations. As noted above, this suggests agency assignment is not "exogenous" and, given the fact that most patients attend the nearest clinic, calls into question any potential instruments for agency assignment. There are also very significant differences in treatment effects across the agencies. Eight of the drift effects, twelve of the completion threshold effects, ten of the dropout probability effects, and eleven of the unit equation effects are significantly different from zero. Differences in the different dimensions of the treatment process across agencies do not appear to be independent. The correlations between the population effect for h_0 and the treatment effects for μ, $\overline{h}$, and λ across the fifteen agencies are reported in Table 7. Agencies whose patient populations are unobservably healthier at admission do not tend to

[41] This table arbitrarily orders agencies. The arbitrary numbering of agencies is necessary to preserve the confidentiality of individual agency information. Hereafter, we identify the individual agencies by this ranking.

[42] Our additive normalization is that the average population effect in the h_{i0} equation is zero (i.e. the j-dimensional row vector α_1 satisfies $\sum_i \alpha_1 \mathbf{A}_i = 0$).

	h_0 (α_1)	μ (δ_1)	$\overline{h}$ (β_1)	λ (γ_1)	$\ln(u)$
average agency	0	−0.0288	6.0802	−3.9834	1.6790
	(0)	(0.0302)	(0.1671)	(0.1881)	(0.1087)
agency 1	−0.3891	0.0554	0.2005	−0.6041	0.2425
	(0.1470)	(0.0206)	(0.2522)	(0.1553)	(0.0513)
agency 2	0.0035	0.0511	2.1703	−0.6601	0.6534
	(0.2093)	(0.0169)	(0.2040)	(0.2427)	(0.1106)
agency 3	0.9086	0.0605	2.1741	0.5162	−0.2726
	(0.1448)	(0.0235)	(0.3285)	(0.1781)	(0.0659)
agency 4	−0.5437	0.0229	−0.7319	−0.4144	0.3000
	(0.1894)	(0.0329)	(0.2176)	(0.2163)	(0.0901)
agency 5	−0.2357	−0.0668	−1.4199	−1.1751	0.3453
	(0.2033)	(0.0251)	(0.2087)	(0.2653)	(0.1011)
agency 6	0.2616	−0.0036	−0.5903	0.2048	−0.1731
	(0.1100)	(0.0106)	(0.2234)	(0.1108)	(0.0403)
agency 7	−0.1775	0.0037	−0.2458	0.3409	−0.2246
	(0.0582)	(0.0089)	(0.0673)	(0.0738)	(0.0331)
agency 8	−0.0643	−0.0050	0.0341	0.0194	0.0665
	(0.0937)	(0.0086)	(0.1973)	(0.0991)	(0.0339)
agency 9	−0.8579	−0.0092	−1.3950	−0.5406	−0.0498
	(0.1908)	(0.0313)	(0.2381)	(0.2463)	(0.1055)
agency 10	−0.8879	−0.0231	−1.6201	−0.8309	0.2514
	(0.1693)	(0.0280)	(0.2943)	(0.2121)	(0.0807)
agency 11	0.8383	−0.0109	1.4175	0.4601	−0.0565
	(0.2207)	(0.0273)	(0.2915)	(0.2429)	(0.0993)
agency 12	−0.2578	−0.0294	−1.3045	−0.0638	−0.0093
	(0.1718)	(0.0159)	(0.3076)	(0.1947)	(0.0777)
agency 13	−0.2549	−0.0331	0.6574	−0.1275	0.2742
	(0.1674)	(0.0138)	(0.2931)	(0.1674)	(0.0694)
agency 14	−0.3125	−0.0446	−0.3767	−0.1406	0.2189
	(0.2180)	(0.0188)	(0.3581)	(0.2017)	(0.0798)
agency 15	0.9714	−0.0525	1.3101	0.2324	0.0693
	(0.1705)	(0.0238)	(0.3278)	(0.2228)	(0.0895)

Table 6: Estimated population (h_0) and treatment effects (in equations for μ, $\overline{h}$, λ) expressed as deviations from the average agency

	h_0	μ	$\overline{h}$	λ	$\ln(u)$
h_0	1	0.088 (0.318)	0.757∗ (0.059)	0.709∗ (0.086)	−0.383∗ (0.1328)
μ		1	0.476 (0.249)	0.135 (0.281)	−0.030 (0.172)
$\overline{h}$			1	0.498∗ (0.103)	0.017 (0.132)
λ				1	−0.746∗ (0.079)
$\ln(u)$					1

Table 7: Correlation between agency population h_0 and treatment effects (μ, $\overline{h}$, λ)

have higher or lower drifts, but do tend to have more demanding completion standards and more trouble keeping their patients until treatment is complete.

Differences in the population and treatment effects can be summarized by comparing the profile of average patients. Table 8 presents the averages of the initial health, the drift of the treatment process, the completion threshold, and the attrition hazard for various patient populations.[43] The requisite health improvement for completion of treatment of the average patient in the whole sample is $\Delta h = \overline{h} - h_0 = 5.3318$. With a drift of $\mu = 0.1722$, this patient would take almost 31 weeks to complete treatment if there was no attrition. However, the patient has a 12.7 percent probability of leaving treatment in the first week, and a 3.7 percent probability of premature exit each week thereafter. Thus, the probability that this patient completes treatment eventually is around 61 percent and the expected

[43]The estimated variance of the Wiener process is $\hat{\sigma}^2 = 0.5065$. The last column of Table 8 reports the average of the time-independent dropout hazard, i.e.:

$$\lambda_i(t) = \exp(\gamma_0 \mathbf{X}_i + (\gamma_1 + \gamma_2 \alpha_1)\mathbf{A}_i + \gamma_2 \varepsilon_i)$$

The actual time-dependent dropout hazard requires multiplying the reported estimate by $e^{\widehat{\gamma}_3 t}$ and making the appropriate adjustment for the first period hazard. However the estimated $\widehat{\gamma}_3$ is close to zero, so the reported estimate is very close to the actual hazard after date 1. The date 1 hazard is higher as discussed earlier.

time to completion conditional on not dropping out is 12.2 weeks.[44] The big difference in expected times to completion shows the importance of controlling for the selection problem in the data. In contrast, the health improvement at completion for the average patient assigned to Agency 1 is $\Delta h = 6.2673$. With a more effective treatment regime indicated by $\mu = 0.2283$, this patient would complete treatment in 27.5 weeks if there was no attrition. The probability of completing treatment eventually is of 62 percent, and the expected time to completion conditional on not dropping out is 14.27 weeks. Finally, Table 9 compares how the average patient in the whole sample fares at each of the agencies. In contrast to Table 8 these differences are due just to treatment effects. The average patient in the population completes treatment at Agency 1 after a health improvement of $\Delta h = 5.5$, expects to complete treatment after 24 weeks, and has a 73.3 percent probability of successful completion. Thus, it appears from these calculations that Agency 1 has a more effective treatment program than the average agency.

All other estimated coefficients are presented and discussed in the Appendix.

4.2 Goodness of fit

How well does the estimated model fit the data? Figures 4 and 5 compare the estimated values and the actual data for time in treatment for different completion status. The predicted drop-out hazards tracks the data very closely. The model fares a bit worse at predicting the timing of the completion of treatment. The estimated model noticeably underpredicts fre-

[44] The probability of completion (conditional on not dropping out) is computed averaging the following probability across patients:

$$\sum_{t=1}^{300} \left(f(t; \widehat{\mu}_i, \widehat{\sigma}_i, \widehat{\Delta h_i})[1 - G(t; \widehat{\lambda}_i(\cdot))] \right).$$

	h_0	μ	$\overline{h}$	$\lambda\ e^{-\gamma_3 t}$
sample	0.6175	0.1722	5.9493	0.0368
Agency 1	−0.0318	0.2283	6.2355	0.0318
Agency 2	−0.8051	0.1279	6.8785	0.0281
Agency 3	1.7953	0.2027	7.6499	0.0286
Agency 4	−0.1404	0.1821	5.4164	0.0457
Agency 5	−0.3045	0.1366	4.4338	0.0091
Agency 6	0.8029	0.1876	4.9980	0.0328
Agency 7	0.6235	0.1922	6.0712	0.0583
Agency 8	0.3661	0.1669	5.8530	0.0409
Agency 9	−0.4575	0.1750	5.0079	0.0531
Agency 10	0.0208	0.1260	5.6259	0.0425
Agency 11	1.9920	0.1801	7.0329	0.0185
Agency 12	0.3885	0.1544	4.7925	0.0420
Agency 13	0.5707	0.1131	7.0552	0.0402
Agency 14	0.1452	0.1380	5.7866	0.0444
Agency 15	1.6040	0.1126	6.2267	0.0146

Table 8: Estimated treatment process for average patients

	h_0	μ	$\overline{h}$	$\lambda e^{-\gamma_3 t}$
Agency 1	0.6175	0.2275	6.1498	0.0201
Agency 2	0.6175	0.2233	8.1196	0.0190
Agency 3	0.6175	0.2327	8.1234	0.0616
Agency 4	0.6175	0.1951	5.2174	0.0243
Agency 5	0.6175	0.1054	4.5294	0.0114
Agency 6	0.6175	0.1685	5.3591	0.0451
Agency 7	0.6175	0.1758	5.7036	0.0517
Agency 8	0.6175	0.1672	5.9835	0.0375
Agency 9	0.6175	0.1630	4.5543	0.0214
Agency 10	0.6175	0.1491	4.3292	0.0160
Agency 11	0.6175	0.1613	7.3669	0.0583
Agency 12	0.6175	0.1428	4.6448	0.0345
Agency 13	0.6175	0.1391	6.6068	0.0324
Agency 14	0.6175	0.1276	5.5726	0.0320
Agency 15	0.6175	0.1196	7.2594	0.0464

Table 9: Estimated average treatment process for the whole sample when treated at different clinics

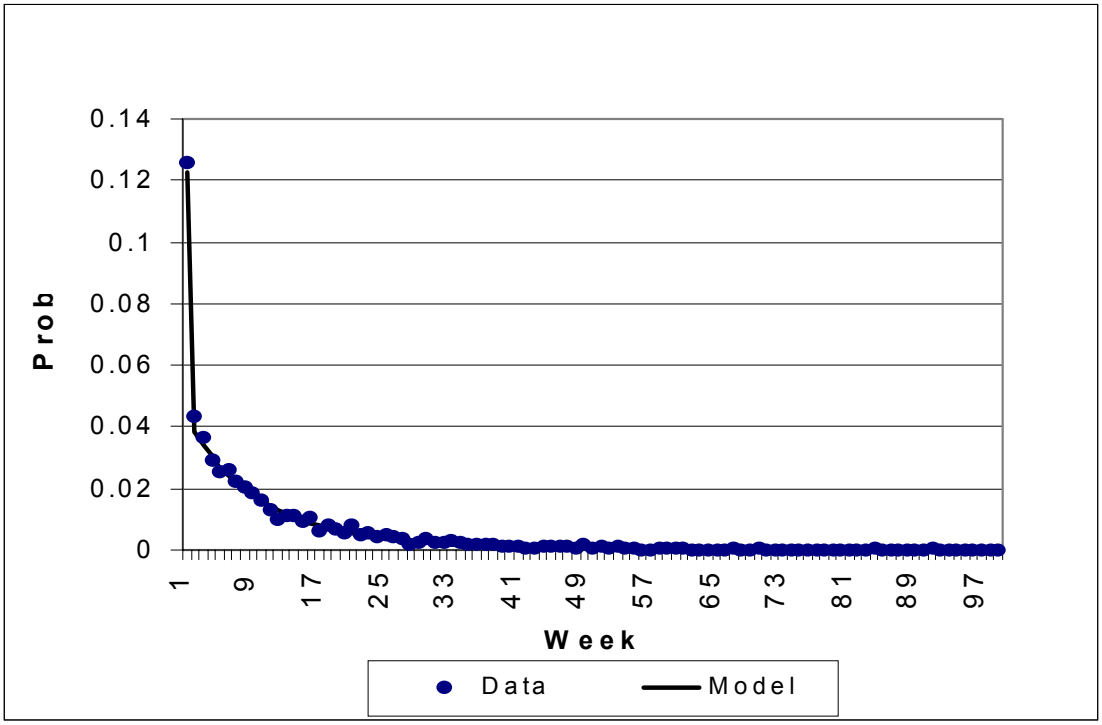

Figure 4: Probability of droping out in date t

quency of completions in weeks 12-15, and slightly overpredicts completions in weeks 21-27. Still, the fits seem reasonable given the parsimony of our model of the treatment process.

We have also estimated the frequency of use of all patients conditional on the probability of completion. A comparison of Tables 3 and 10 shows that the model predicts quite well the frequency of use at completion of patients who are light alcohol users at admission, while it underpredicts the probability of abstinence at completion of heavier drinkers at admission.

4.3 Productivity comparisons

We next compare the productivities of these fifteen treatment agencies. Our estimated model enables us to compute the hypothetical aggregate health improvement and total costs for *any*

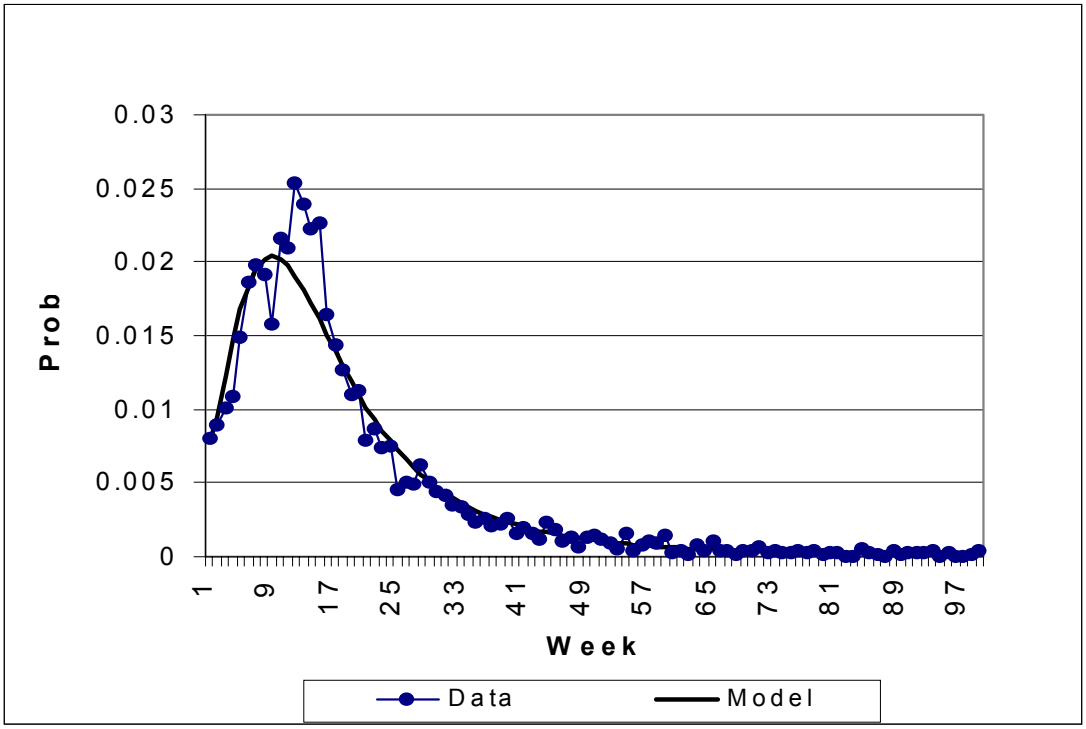

Figure 5: Probability of Completing in date t

population of patients assigned to *any* treatment program. We consider sixteen possible patient populations – the sample populations of each of the fifteen treatment agencies as well as the entire sample population, and can match these sixteen patient populations to the fifteen agencies in all possible ways. To better visualize these conceptual experiments consider the problem of comparing the productivity of two different teachers in a school. A simple and natural experiment for comparing teaching productivities would be to switch teachers,

Frequency of use at:		Expected frequency of use at completion			
		abstinent	$<$ once daily	$\geq$ once daily	Total
	abstinent	0.9864	0.0135	0.00015	1.0
admission	$<$ once daily	0.9236	0.0742	0.00223	1.0
	$\geq$ once daily	0.7564	0.2252	0.01844	1.0

Table 10: Average estimated frequency of use across all patients conditional on the probability of completion

keeping classes fixed, and measure the performance of each class under each teacher. This type of comparison implicitly controls for different characteristics of the two classes. Our simulations emulate similar controlled experiments in a health services treatment setting.

Productivity is a standard measure of economic efficiency. In manufacturing industries, productivity is measured simply by dividing aggregate output by aggregate cost. In a health services setting, however, the measurement of productivity is complicated by the fact that it is not clear how to most appropriately measure output. The most obvious output measure in our setting is health improvement, i.e. the change in the latent variable representing the health of the patient. Thus, output is $\Delta h_i = \overline{h}_i - h_{0i}$ for a patient who completes treatment (since by definition, patients who have completed treatment have achieved level $\overline{h}_i$). For patients who drop out prematurely, output is $h_{ti} - h_{0i}$, where t is the time of dropout. Aggregate output is measured by summing expected health improvement over the patient population being considered. Aggregate cost is measured by multiplying an agency's estimated total units of treatment by its unit cost of treatment. These measures enable our productivity calculations.[45]

More precisely, consider the expected total increase in latent health that would result if agency k were to treat population j. The fact that we have separated population and treatment effects is critical for this procedure. When agency k treats population j, we use the treatment effects of agency k, but the population effects of population j. In other words, the fact that population j is relatively hard or easy to treat *goes with* that population when we move the population to different agencies. The fact that agency k is more or less effective

[45] The Appendix considers aggregate abstinence as an alternative output measure. In the the next section, we consider some conceptual problems with using these productivity estimates to measure economic efficiency, and adopt a less restrictive approach.

(δ_1), or has different completion thresholds (β_1), etc. *stays with* the agency as we assign to it different patient populations. In some cases, we compare our results and policy predictions to simulations in which we naively assume that all the patient populations are unobservably identical, i.e. where there are no population effects.

Given our estimated parameters $(\widehat{h}_{i0}, \widehat{\mu}_{ik}, \widehat{\overline{h}}_{ik}, \widehat{\lambda}_{ik}, \hat{\sigma}; i = 1, ...N)$, the total expected increase in health of population j at agency k is given by the following formula:

$$\sum_{i=1}^{N} A_{ij} E_{\varepsilon_i}[(\widehat{\overline{h}}_{ik} - \widehat{h}_{i0}) \sum_{t=1}^{\infty} f(t; \widehat{\mu}_{ik}, \widehat{\sigma}, \widehat{\overline{h}}_{ik} - \widehat{h}_{i0}) \left(1 - G(t; \widehat{\lambda}_{ik})\right) dt +$$

$$\sum_{t=1}^{\infty} \left(\widehat{\mu}_{ik} t - 2(\widehat{\overline{h}}_{ik} - \widehat{h}_{i0}) \exp\left(\frac{2\widehat{\mu}_{ik}(\widehat{\overline{h}}_{ik} - \widehat{h}_{i0})}{\widehat{\sigma}^2}\right) \frac{\Phi\left(\frac{-(\widehat{\overline{h}}_{ik} - \widehat{h}_{i0}) - \widehat{\mu}_{ik} t}{\widehat{\sigma}\sqrt{t}}\right)}{1 - F(t, \widehat{\mu}_i, \widehat{\sigma}, \widehat{\overline{h}}_{ik} - \widehat{h}_{i0})} \right)$$

$$\times g(t; \widehat{\lambda}_{ik})(1 - F(t, \widehat{\mu}_i, \widehat{\sigma}, \widehat{\overline{h}}_{ik} - \widehat{h}_{i0}) dt] \tag{25}$$

The first line of the expression gives the expected health improvement conditional on completion multiplied by the probability of completion. The next two lines give the expected health improvement conditional on premature exit at each date t, multiplied by the probability of exit at that date, and summed over all possible exit dates.[46] To actually compute (25), we simulate the expectation over the distribution of the unobserved characteristics ε_i.[47]

We divide this measure of aggregate output by aggregate cost to calculate productivity. Aggregate cost is measured by multiplying the predicted treatment units for patient i in agency k, from equation (22), by the unit cost of agency k. As explained earlier, we translate all units of treatment into equivalent units of individual therapy, and use the unit cost of

[46] The formula for expected health improvement conditional on not having completed at date t can be found in Lancaster (1990, pp. 119-121).

[47] The simulation truncates the infinite sum of completion dates at $T = 300$. As the probability of a patient staying more than even 200 weeks is extremely small, there is no apparent differences between truncating the sum at 200 or 300.

individual therapy for fiscal year 1993. In doing so, we are implicitly assuming that unit costs are exogenously fixed and do not vary with the patient population. The reason for choosing fiscal year 1993 for these productivity comparisons is that all agencies had programs running in that year and therefore we have comparable cost information.

The estimated productivities for agencies treating their own patient population and the entire sample population are presented in Table 11. There are three columns in the table.[48,49] The first reports the productivities of each agency when treating its own population. These productivities do not control for differences in either observed or unobserved patient characteristics across agencies. The second column simulates productivities for each agency treating the entire sample population under the naive assumption that there are no population effects, i.e. that the estimated agency fixed effects in the drift, threshold, dropout, and units equations are entirely treatment effects. This comparison controls for observable patient characteristics (because all agencies are treating the same population) but makes no attempt to control for population differences in the unobservable characteristic. The last column reports simulated productivities of the entire population that account appropriately for the estimated population effects. This last simulation accounts for both observable and unobservable differences across patient populations.

[48] In making these calculation we assume that all patients enter treatment beginning in January 1993. We adjust the time dependent variables in our estimated treatment model accordingly, and use unit cost data for this year. The numbers presented in Tables 11, 17, and 18 (see appendix) have been multiplied by 1000.

[49] As the estimated productivities are highly non-linear functions of the parameters, we compute their standard errors by bootstrapping the data and estimation. Since actual estimation is very time consuming (1 to 2 weeks of CPU time), we follow Andrews(1999) suggestion and use a K-step bootstrap. Using the actual estimates as starting parameters, we take K Gauss-Newton steps with the resampled data. While Andrews suggests that $K = 10$ or 20 works reasonably in many cases, we use $K = 100$, which still leads to a significant time savings.

The same bootstrapping methodology is used in our technology transfer experiment below. This is necessary because we employ separate estimation procedures for the basic model and one for the units equation. The bootstrap simulations provide confidence intervals that account for the correlations between the estimated parameters in the two sets of equations.

Table 11 reveals significant differences of the productivities of the agencies. Agency productivities are as high as 12.87 and as low as 5.62, a difference of two to three times the "health" per dollar spent. There are also notable differences in the ranking of agencies across the columns. Ignoring both observable and unobservable differences in the agency populations leads to the conclusion that Agency 1 is the top performer and Agency 13 the worst. In contrast, recognizing and controlling only for observable population differences shows that Agency 2 is the best performer and Agency 15 the worst, while controlling for all differences in population, Agency 2 remains the best and Agency 10 is the worst.

The difference between the estimated productivites in first and second columns of Table 11 can be interpreted as capturing the effect of controlling for the observable characteristics. Analogously, the difference between the second and third columns captures the effect of controlling for unobservable characteristics. The variance of the first effect (the difference between columns 1 and 2) across agencies is 1.52 while the variance of the second effect (the difference between columns 2 and 3) is 2.00. This suggests that controlling for differences in unobserved characteristics across agencies is as much, if not more, important than controlling for observed characteristics.

4.4 Efficiency frontiers

While productivities often are convenient and intuitive measures of economic efficiency, productivity comparisons can tell a misleading story in a health services setting. While manufacturing industries do not produce outputs without inputs, it is not necessarily true that patients would not recover without health treatment. Unfortunately, our data set does not contain any observations on patients *not* entering treatment, which prevents us from directly

	Own population		Entire population NOT accounting for pop. eff.		Entire population accounting for pop. eff.	
Agency	product.	ranking	product.	ranking	product.	ranking
1	12.87 (0.895)	1	12.88 (0.940)	2	10.96 (0.842)	2
2	12.08 (1.480)	2	15.50 (1.477)	1	14.95 (1.347)	1
3	8.38 (0.591)	11	8.11 (0.620)	13	9.33 (0.902)	4
4	9.45 (1.517)	6	10.48 (1.359)	3	8.36 (1.045)	7
5	10.94 (1.450)	3	10.01 (1.619)	4	8.45 (1.219)	6
6	9.10 (0.654)	8	8.37 (0.702)	11	8.71 (0.894)	5
7	9.34 (0.587)	7	8.86 (0.645)	7	8.09 (0.542)	9
8	7.90 (0.528)	12	8.04 (0.553)	14	7.55 (0.542)	12
9	8.83 (1.144)	10	8.97 (1.109)	6	6.27 (0.846)	14
10	7.12 (0.733)	13	8.55 (0.719)	10	5.67 (0.750)	15
11	10.30 (1.245)	4	8.74 (1.216)	8	10.17 (1.560)	3
12	9.96 (0.937)	5	9.58 (0.930)	5	8.17 (0.938)	8
13	5.62 (0.814)	15	7.26 (0.758)	12	6.61 (0.685)	13
14	8.87 (1.228)	9	8.70 (1.208)	9	7.59 (1.048)	11
15	6.69 (0.782)	14	6.31 (0.803)	15	7.87 (1.319)	10

Table 11: Agency rankings according to estimated productivities in terms of health

addressing this issue.[50]

Using the productivity measures as comparative efficiency indicators implicitly assumes that patients would not change health status if they were not to receive treatment. Figure 6, which graphs total costs vs. total health improvement, illustrates the problem. Given a particular patient population, we can plot different agencies on this graph according to their total costs of treating this population and the total health improvement achieved. Productivities equal the slope of the line between each agency point and the origin. Thus in Figure 6 Agency A has the highest productivity and Agency C the lowest. Suppose, however, that instead of getting 0 health improvement with no treatment, untreated patients were to achieve a total health improvement of 10. In this case, the "true" productivity of treatment would equal the slope of the line connecting each agency point to the point 10 on the vertical axis, and the ranking of productivities would be exactly reversed.

Another possible problem with a productivity measure based on the ratio of total output per total cost is the possibility of diminishing or increasing returns to treatment at the level of the individual patient. Suppose some agencies tend to give each patient more intense treatment than other agencies (e.g. more units of treatment per week), one might expect to find diminishing returns, i.e. patients marginal health improvement from an extra dollar of treatment declines as the intensity of treatment increases. As can be easily observed from the example in figure 6, the productivity measures are simply the magnitude of the slopes of straight lines joining each point A, B, and C with the origin, and, therefore, in the case where the true production function is concave we may wrongly conclude that agency C is performing badly when in fact agency C is on the production frontier.[51]

[50] See the Appendix for some observations in reference to the received substance abuse treatment literature.

[51] Note that our model implicitly does not allow increasing or decreasing returns to the amount of time a

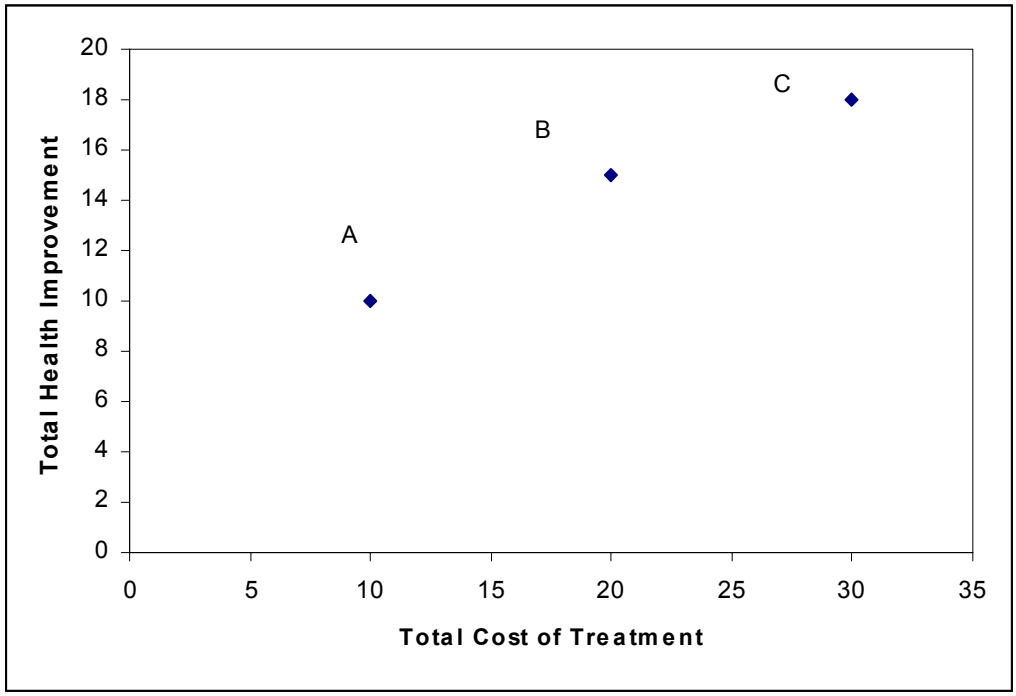

Figure 6: An example of health improvement versus total costs

There are several possible avenues of attack on these problems. One possibility is to use our structural model to "scale up" or "scale down" production at the various agencies to bring them to comparable points. For example, we could scale up production at Agency A and scale down production at Agency C so that all three agencies produce the same level of health. One could then compare costs to rank the different agencies. While this is possible in our model, it is not clear on what margins to scale production. For example, for Agency C we could either decrease its completion standards or decrease the amount of treatment it is giving patients per week. Another issue is that in doing this, we would need to presume that the other treatment effects for the agency remain the same. It might be inappropriate, for example, to assume that increasing the amount of treatment per week doesn't affect

patient spends in treatment or scale economies in the number of patients receiving treatment.

dropout probabilities. A second solution would be to postulate social preferences over health/cost space, e.g. draw indifference curves in Figure 6 and rank agencies based on these indifference curves. However, we have no particular insight into how society values latent health improvements.

We adopt what seems like a less presumptuous approach by limiting ourselves to conclusions about the relative performance of agencies that have a dominance relationship. That is, if a certain agency provides a given population with no more health improvement than would another agency and yet has higher total costs, then we conclude that the first agency is dominated and not performing efficiently. On the one hand, this relative performance comparison makes no assumptions about either the effects of no treatment, economies of scale in treatment intensity, or societal preferences, other than that the same amount of treatment for lower cost is better. On the other hand, this is admittedly a conservative approach to comparing the performances of the treatment agencies, because it implicitly only values cost reductions and not health improvements.

Figure 7 plots our fifteen agencies in health/cost space. These points examine the situation in which each agency treats the entire population, controlling appropriately for population and treatment effects. An agency is dominated in this experiment if there is another agency to the northeast of it on the diagram. As can be seen, there are a number of dominated agencies. For example, Agencies 9 and 10 provide less health than Agency 11, but spend almost double the money on treatment; Agency 13 provides less health than Agency 4, but spends $500,000 more. The only agencies that are not dominated are Agencies 2, 4, 6, 11, and 12. One can visualize a downward sloping efficiency frontier on the diagram, specifying for any given total cost the total health improvement that an efficient agency can

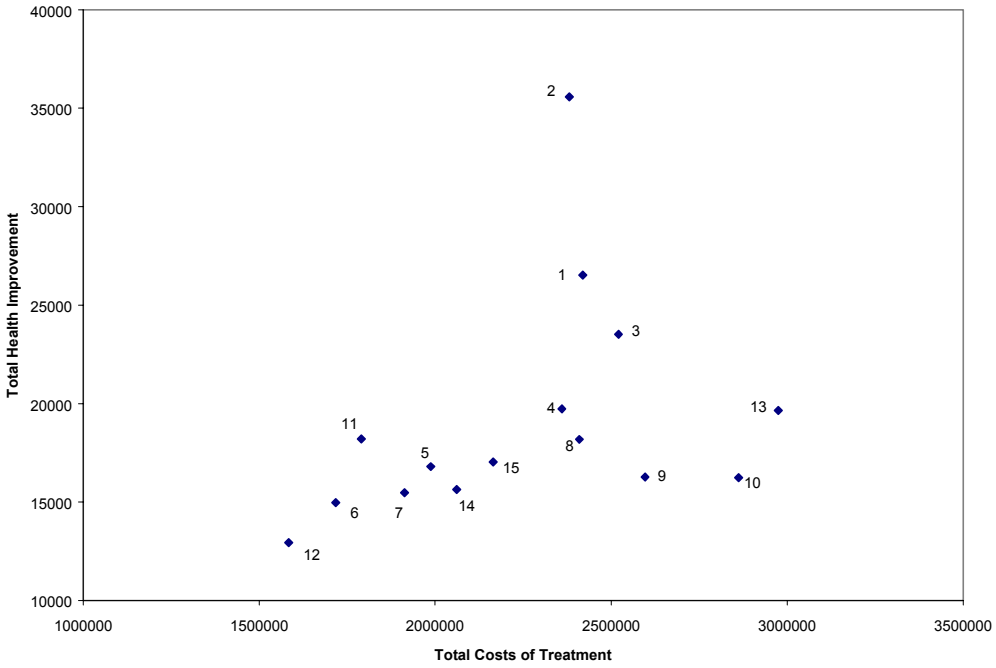

Figure 7: Health improvement versus costs

produce. We can conclude that the dominated agencies are not on this frontier, i.e. would not provide treatment to the entire population efficiently. While we cannot necessarily conclude that the undominated firms are on the true frontier (i.e. the frontier may lie to the northeast), we will loosely describe them as firms on the efficiency frontier.

These efficiency figures are just point estimates. Using our resampling bootstrap methods (see footnote 46), we can statistically test whether firms are inefficient. For a given number of bootstrapped parameter vectors, we compute how many times a particular agency is undominated, i.e. when the agency is on the efficiency frontier. The proportion of bootstrap repetitions when this occurs serves as a p-value for the null hypotheses that a firm is on the frontier. For example, if an agency is on the frontier in only 3% of the bootstrap repetitions, we reject the null hypothesis that the agency is efficient with 97% confidence. We can reject

that Agencies 8, 9, 10, and 13 are efficient with at least 95% confidence. Agency 2 is the only firm that we can reject being dominated by another firm.[52]

4.5 Transfering of best practices

The above analysis considers the health/cost trade-off for the entire patient population. We also examine these trade-offs for each particular agency's patient populations by constructing similar figures. For each agency's population we compare the estimated outcomes and costs of treatment against the simulated outcomes and costs if the population were treated by other agencies. In particular, we note whether there is another agency that can provide at least the same amount of health as the originating agency, but at lower cost. If there are multiple agencies that can provide more treatment at less cost, we choose the lowest cost agency. The final goal of this exercise is to estimate a lower bound of how much money Maine could save (without compromising health outcomes) by transferring the practice styles of the more efficient agencies.

Table 12 reports estimated savings that could be achieved by "transferring best practices" of more efficient agencies. We order the results by the total cost savings, i.e. transfers that save the most money are listed first. The first transfer has Agency 10 transferring its practice style to Agency 5, saving $46,000, more than 2% of the overall expenditures (at all fifteen clinics). There are a total of ten cost-saving transfers - there is no potential cost savings for Agencies 2, 4, 6, 11, and 12. The total cost saving due to the ten transfers is $163,000, 7.4% of the total budget for the fifteen agencies. Bootstrapped standard errors show that

[52] More details on these tests are in the Appendix in comparison to similar tests using abstinence as an output measure.

transferring practice styles between clinics can lead to significant cost savings.[53,54]

The virtue of using our measure of latent health as an output measure is that it effectively combines information from multiple output measures. However, one can also use single, yet more tangible, output measures. Table 1 in the introduction reproduces the analysis using abstinence. Other output measures such as, for instance, completion of treatment could be used as well. Comparing tables 1 and 12, the agencies that would most benefit from technology transfer are similar, but the agencies who transfer the practices are somewhat different. As noted above, we favor the "latent health" results since they use information from all the output measures in our dataset.

4.6 Explaining superior performance

In this section we present some evidence that supports the idea that the superior performance of certain agencies is idiosyncratic, i.e. due to differences in practice styles, and not related to observable and exogenous agency characteristics. We concentrate on two agency characteristics: size, and importance of group treatment vis-a-vis individual therapy. Size is measured by the average annual budget for (individual, group and family) outpatient treatment,[55] and group units is the ratio between adjusted units of group treatment and total units of treatment.[56] We normalized size by dividing it by its standard deviation.

[53] Importantly, with our bootstrapped parameter vectors, we do not recompute the optimal transfers. As such, some of these transfers can potentially cost money (for particular bootstrapped parameter vectors).

[54] Note that while the direct effect of these transfers is to save money, there are also a potential side effect of improving health outcomes (since in some cases the new "owners" not only save money but produce more health). For both outcome variables, health improvement (abstinence rates) goes up a total of about 1-2% with the transfers.

[55] We ignore the budget for evaluation because we have excluded from our sample patients who only received evaluation.

[56] We use the same approach as in Section 2.2. In order to obtain a total number of units we add family and group units after we multiply them by their weights. Again their weights are given by their relative costs in terms of "individual" treatment. Lastly evaluation units are assumed to be equivalent to "individual"

Controlling for Unobservable Population Effects				
Population of Agency	Treated by Agency	$ Savings	Cumulative $ Savings	Cumulative % Savings
10	5	46131.43	46131.43 (14573.13)	0.0210 (0.0067)
13	4	30948.07	77079.50 (22078.73)	0.0351 (0.0100)
15	11	18765.26	95844.76 (24922.34)	0.0436 (0.0114)
7	11	17265.06	113109.82 (44368.38)	0.0515 (0.0202)
3	2	16893.51	130003.33 (52250.61)	0.0592 (0.0237)
14	11	9441.98	139445.31 (55455.86)	0.0635 (0.0252)
8	4	8958.95	148404.26 (72829.64)	0.0675 (0.0331)
9	5	8676.63	157080.89 (73558.75)	0.0715 (0.0335)
5	11	3253.42	160334.31 (74488.06)	0.0730 (0.0339)
1	2	2217.06	162551.37 (81150.72)	0.0740 (0.0369)
11	11	0	162551.37 (81150.72)	0.0740 (0.0369)
6	6	0	162551.37 (81150.72)	0.0740 (0.0369)
4	4	0	162551.37 (81150.72)	0.0740 (0.0369)
12	12	0	162551.37 (81150.72)	0.0740 (0.0369)
2	2	0	162551.37 (81150.72)	0.0740 (0.0369)

Table 12: Cost reducing technology transfers, controlling for population effects in terms of health

agency category	good		medium		poor	
	mean	s.d.	mean	s.d.	mean	s.d.
size	0.037	0.032	0.066	0.063	0.067	0.055
% group	1.136	0.555	1.663	1.261	1.291	1.261

Table 13: Mean and variances of agency characteristics by agency category

The correlation between these two agency characteristics is around 0.35. In order to show approximately how these agency characteristics correlate with performance, we categorize agencies into three groups according to their proximity to the efficiency frontier (see figure 7).[57] Table 13 shows the mean and variance for the variables across the three agency categories.[58] These results suggest that superior performance is likely to reflect different "practice styles."

5 Conclusions

We have developed and applied a methodology for comparing the performance of providers of alcohol abuse treatment in Maine. Our results should be interpreted tentatively for drawing specific policy conclusions about alcohol abuse treatment in Maine. It is certainly possible to analyze more data more extensively. Nevertheless, our analysis of the Maine data illustrates a sound methodology for relative performance evaluations that is potentially a useful public policy tool. The methodology enables a separate identification of agency "popula-

treatment units.

[57] Good performers include agencies {2, 4, 6, 11, 12}, bad performers include agencies {8, 9, 10, 13} and medium performers include agencies {1, 3, 5, 7, 14, 15}.

[58] We made simple tests on the equality of means across agency categories. We were not able to reject that the mean of "percentage group" of good performers is lower than the mean for medium and bad performers. For the variable size we were never able to reject the null of common means across agency categories. All tests were done assuming the variables were normally distributed. Finally we ran simple OLS regressions with the classification into the three groups as the dependent variable and size and percentage group as explanatory variables. Conditioning on percentage group made the size coefficient positive although small and not statistically different from zero. The group coefficient was negative but not statistically significantly different from zero.

tion effects" and "treatment effects" which enables us to compare how different agencies would perform treating the same populations. We conclude that different agencies have significantly different treatment effects, governing the effectiveness of treatment, standards of completion of treatment, and the ability to retain patients until they complete treatment. To the extent that these differences reflect differences in practice style or management techniques, we conclude that Maine potentially could improve the cost effectiveness of publicly funded treatment substantially by identifying and transferring best practices.[59]

References

[1] Aitken, L.S. (1986), "Retrospective Self-Reports by Clinets Differ from Ongoing Reports: implications for the evaluation of drug treatment programs," International Journal of Addictions, 21(7), 767-788.

[2] Andrews, Donald W. K.(1999), "Higher-order Improvements of a Computationally Attractive k-step Bootstrap for Extremum Estimators," *Cowles Foundation Discussion Paper,* No *1230.*

[3] Anglin, M. D. and Y. Hser (1990), "Treatment of Drug Abuse," in: Tonry, M. and J.Q. Wilson (Eds), *Crime and Justice: an annual review of research*, vol 13. University of chicago Press.

[4] Apsler, Robert and Wayne Harding (1991), "Cost-Effectiveness Analysis of Drug Abuse Treatment: current status and recommendations for future research," in: *Drug Abuse Services Research Series,* No 1: Background Papers on Drug Abuse Financing and Ser-

[59] There is a large literature on the effectiveness of substance abuse treatment. The Appendix contains a brief discussion of how our approach and results relate to that literature.

vices Research. National Institute on Drug Abuse. DHHS Pub. No. (ADM) 91-17777. Washington, D.C., 58-81.

[5] Apsler, Robert (1991), "Evaluating the Cost-Effectiveness of Drug Abuse Treatment Services," in: *Economic Costs, Cost-Effectiveness, Financing and Community-Based Drug Treatment*. Rockville, MD: National Institute on Drug Abuse, Research Monograph Series 113, 17-26.

[6] Ball, John C. and Alan Ross (1991), "The Effectiveness of Methadone Maintenance Treatment: patients, programs, services, and outcome." Springer-Verlag New York Inc.

[7] Bound, John (1991),"Self-Reported vs Objective measures of Health in retirement models," *Journal of Human Resources,* 26, 107-137.

[8] Butler, J. S., Richard V. Burkhauser, Jean Mitchell and Theodore P. Pincus (1987), "Measurement Error in Self-Reported Health Variables," *The Review of Economics and Statistics*, 69, 644-650.

[9] Commons, Margaret, Thomas G. McGuire, and Michael H. Riordan (1997), "Performance Contracting for Substance Abuse Treatment," *Health Sciences Research*, 32 (5), 631-650.

[10] Commons, Margaret and Thomas G. McGuire (1997), "Some Economics of Performance Based Contracting for Substance Abuse Services," in: Egertson, Joel A, Daniel M. Fox, and Alan I. Leshner (Eds), *Treating Drug abusers Effectively*. Milbank Memorial Fund and Blackwell Publishers, 223-249.

[11] Drummond, D. Colin (1997), "What I Would Most Like to Know. Alcohol Interventions: do the best things come in small packages?," *Addiction,* 92(4), 375-379.

[12] Eckstein, Zvi and Kenneth Wolpin (1999), "Why Youths Dropout of High School: the impact of preferences, opportunities and abilities," *Econometrica,* 67(6), 1295-1339.

[13] Emrick, Chad D. (1975), "A Review of Psychological Oriented Treatment for Alcoholism II - the relative effectiveness of different treatment approaches and the effectiveness of treatment versus non-treatment," *Journal of Studies on Alcohol*, 36, No 1, 88-108.

[14] Finney, John W., Annette C. Hahn and Rudolf H. Moos (1996), "The Effectiveness of Inpatient and Outpatient Treatment for Alcohol Abuse: the need to focus on mediators and moderators of setting effects," *Addiction,* 91(12), 1773-1796

[15] French, Michael T., Gary A. Zarkin, Robert L. Hubbard, and J. Valley Rachal (1991), "The Impact of Time in Treatment on the Employment and Earnings of Drug Abusers," *American Journal of Public Health*, 81(7), 904-907.

[16] Gerstein, D. R. and H. J. Hardwood, (Eds) (1990), *Treating Drug Problems.* Washington, D.C.: Institute of Medicine, National Academy Press.

[17] Grant, Kathryn A., Lisa T. Arciniega, J. Scott Tonigan, William R. Miller and Robert Meyers (1997), "Are Reconstructed Self-Reports of Drinking Reliable?," *Addiction,* 92(5), 601-606.

[18] Heckman, J. J. (1997), "Instrumental Variables: a Study of Implicit Behavioral Assumptions Used in Making Program Evaluations", *Journal of Human Resources*, 32, 442-461.

[19] Heckman, J. J. and Jeffrey A. Smith (1995), "Assessing the Case for Social Experiments," Journal of Economic Perspectives, 9(2), 85-110.

[20] Higgins, Stephen T., Alan J. Budney, Warren K. Bickel, and Gary J. Badger (1994), "Participation of Significant Others in Outpatient Behavioral Treatment Predicts Greater Cocaine Abstinence," *American Journal of Drug and Alcohol Abuse*, 20(1), 47-56.

[21] Høyer, Georg, Odd Nilssen, Tormod Brenn, and Helge Schirmer (1995), "The Svalbard Study 1988-89: a unique setting for validation of self-reported alcohol consumption," *Addiction*, 90, 539-544.

[22] Imbens, Guido W. and Joshua D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62 (2), 467-475.

[23] Jaffe, Jerome, H. (1984), "Evaluating Drug Abuse Treatment: a comment on the state of the art," in: Tims, F. M. and J.P. Ludford (Eds), *Drug Abuse Treatment Evaluation: Strategies, Progress and Prospects.* National Institute on Drug abuse Research Monograph No. 51. Rockville, MD, 13-28.

[24] Joe, George W., D. Dwayne Simpson and S.B. Sells (1994), "Treatment Process and Relapse to Opiod Use during Methadone Maintenance," *American Journal of Drug and Alcohol Abuse*, 20(2), 173-197.

[25] Keane, Michael and Kenneth Wolpin (1997), "The Career Decisions of Young Men," *Journal of Political Economy,* 105, 3rd issue, 473-522.

[26] Lancaster, Tony (1990), *The Econometric Analysis of Transition Data*, Econometric Society Monographs No 17..Cambridge University Press.

[27] Long, Clive G., Martin Williams, and Clive R. Hollin (1998), "Treating Alcohol Problems: a study of programme effectiveness and cost-effectiveness according to length and delivery of treatment," *Addiction,* 93(4) 561-571.

[28] Longabaugh, Richard, Philip W. Wirtz, Allen Zweben, and Robert L. Stout (1998), "Network Support for Drinking, Alcoholics Anonymous and Long-Term Matching Effects," *Addiction,* 93(9), 1313-1333.

[29] Longabaugh, Richard, Martha Beattie, Nora Noel, Robert Stout, and Paul Malloy (1993), "The Effect of Social Investment on Treatment Outcome," *Journal of Studies on Alcohol,* 54(4), 465-478.

[30] Longabaugh, Richard, Barbara McCrady, Ed Fink, Robert Stout, Thomas McAuley, Cathy Doyle and Dwight McNeill (1983), "Cost-Effectiveness of Alcoholism Treatment in Partial vs Inpatient Settings - six-month outcomes," *Journal of Studies on Alcohol,* 44, No 6, 1049-1071.

[31] Lu, Mingshan and Thomas G. McGuire (2001), "The Productivity of Outpatient Treatment for Substance Abuse," *Journal of Human Resources,* forthcoming.

[32] Lu, Mingshan (1999), "Separating the "True Effect" from "Gaming" in Incentive Based Contracts in Health Care," *Journal of Economics and Management Strategy,* 8, issue 3, 383-431.

[33] Machado, Matilde P. (2001), "Dollars and Performance: Treating Alcohol Misuse in Maine," *Journal of Health Economics*, 20 (4), 645-672.

[34] Machado, Matilde P. (1998), "Dollars and Performance: Cost-effectiveness of the Public Provision of Substance Abuse Treatment in Maine," Boston University Dissertation.

[35] Maitso, S. A., J. R. McKay, and G. J. Connors (1990), "Self-Report Issues in substance Abuse: state of the art and future directions," Behavioral Assessment, 12, 117-134.

[36] Manski, Charles F. (1996), "Learning about Treatment Effects from Experiments with Random Assignment of Treatments," *Journal of Human Resources*, 31 (4), 709-733.

[37] McKay, James R. (1996), "Comments on Finney et al. "The Effectiveness of Inpatient and Outpatient Treatment for Alcohol Abuse: the need to focus on mediators and moderators of setting effects," *Addiction,* 91(12), 1803-1820.

[38] McLellan, Thomas A., Teresa A. Hagan, Marvin Levine, Frank Gould, Kathleen Meyers, Mark Bencivengo, and Jack Durell (1998), "Supplemental Social Services Improve Outcomes in Public Addiction Treatment," *Addiction,* 93(10), 1489-1499.

[39] McLellan, A. Thomas, George E. Woody, David S. Metzger, Jim McKay, Jack Durell, Arthur I. Alterman, Charles P. O'Brien (1997), "Evaluating the Effectiveness of Addiction Treatments: reasonable expectations and appropriate comparisons," in: Egertson, Joel A, Daniel M. Fox, and Alan I. Leshner (Eds), *Treating Drug Abusers Effectively.* Milbank Memorial Fund and Blackwell Publishers 7-40.

[40] McLellan, A. Thomas, Grant R. Grissom, Peter Brill, Jack Durell, David S. Metzger, Charles P. O'Brien (1993), "Private Substance Abuse Treatments: are some programs more effective than others?," *Journal of Substance abuse Treatment,* 10, 243-254.

[41] McLellan, A. Thomas, Isabelle O. Arndt, David S. Metzger, George E. Woody, Charles P. O'Brien (1993), "The Effects of Psychosocial Services in Substance Abuse Treatment," *Journal of the American Medical Association*, 269 (15), 1953-1959.

[42] McLellan, A. Thomas, L. Luborsky, C. P. O'Brien, and G. E. Woody (1980), "An Improved Evaluation Instrument for Substance Abuse Patients: the Addiction Severity Index," *Journal of Nervous and Mental Disorders,* 168, 26-33.

[43] Moos, Rudolph, John Finney, and Ruth C. Cronkite (1990), *Alcoholism Treatment.* Oxford University Press.

[44] Olley, Steven and Ariel Pakes (1996), "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica,* 64(6), 1263-1298.

[45] Roche, Ann M., Michael D. Parle, Joanne M. Stubbs, Wayne Hall and John Saunders (1995), "Management and treatment Efficacy of Drug and Alcohol Problems: what do doctors believe?," *Addiction,* 90, 1357-1366.

[46] Shen, Yujing (1998), "Selection Incentives in a Performance-Based Contracting System," Boston University manuscript.

[47] Strohmetz, D. B., Alterman, A. I. and Walter, D. (1990), "Subject Selection Bias in Alcoholics Volunteering for a Treatment Study," *Alcoholism*: *Clinical and Experimental Research,*14, 736-738.

[48] Tims, Frank M., Bennett W. Fletcher, and Robert L. Hubbard (1991), "Treatment Outcomes for Drug Abuse Clients," in: Pickens, C.G., C. G. Leukefeld, and C. R. Schuster (Eds), *National Institute on Drug Abuse Research Monograph 106.* Rockville, MD: US Department of Health and Human Services.

[49] Walsh, Diana Chapman, Ralph Hingson, Daniel Merrigan, Suzette Morelock Levenson, Adrienne Cupples, Timothy Heeren, Gerald Coffman, Charles Becker, Thomas Barker,

Susan Hamilton, Thomas G. McGuire, Cecil Kelly (1991), "A Randomized Trial of Treatment Options for Alcohol-Abusing Workers," *The New England Journal of Medicine,* September 12, 325 (11), 775-782.

Frequency of use parameters			
cutoffs		error terms	
lower	upper	standard deviation	correlation
−3.4309	0.7787	3.0049	0.5781
(0.1436)	(0.1045)	(0.0915)	(0.0304)

Table 14: Estimated frequency of use model

A Appendix

A.1 Other estimated parameters

Our estimated model contains a large number of coefficients in addition to the estimated treatment and population effects discussed in the text. We begin with the frequency of use model, i.e. the ordered probit model relating current health and frequency of alcohol use. Recall that we have defined a categorical variable taking on three possible values to describe frequency of use. Thus, we estimate two cut-off values. We also estimate the joint distribution of the error terms.[60] The estimates are presented in Table 14.

According to the estimates, a patient is using alcohol at least once a day if $h_{i0} + \eta_i < -3.4309$, and is abstinent if $h_{i0} + \eta_i > 0.7787$. A similar interpretation applies for frequency of use upon completion, for which the relevant latent variable is $\bar{h}_i + \bar{\eta}_i$. The standard errors of the estimates are in parentheses, and indicate that our estimates for this frequency of use model are reasonably precise. The relatively large variances on the error terms (relative to the distance between the cutoffs) indicate significant heterogeneity in the population of patients regarding the relationship between health status and alcohol use. The fact that much of this variance is idiosyncratic (i.e. the correlation coefficient between the error term at the start of treatment and the error term at the

[60] The coefficient on h is set to equal to unity. This normalization defines the units of our latent variable determining frequency.

end of treatment is only 0.58) suggests that the variable "reduction in frequency of use" may not, by itself, be a very precise outcome measure for patients.

We next turn to the effect of patient characteristics (X) on the treatment process. These parameter estimates are collected in Table 15. Standard errors are in parentheses. The columns present the estimated coefficients on X_i in 1) initial health (h_0), 2) the drift of the treatment process (μ), 3) health at completion ($\overline{h}$), 4) the probability of dropping out of treatment (λ), and 5) the units equation ($\ln(u)$). These coefficients require cautious interpretations, as they incorporate both the direct effect of the observable characteristic and an indirect effect through the conditional mean of the unobservable characteristic. The two effects are not identified separately. Recognizing this ambiguity, the estimates have straightforward interpretations. For example, positive coefficients in the μ equation indicate a better response to treatment. Positive coefficients in the equations for h_0 and $\overline{h}$ also indicate that the patient consumes alcohol less frequently at admission and completion. Positive coefficients in the equation for $\overline{h}$ and negatives coefficients in the equations for h_0, μ, and λ indicate that the patient is likely to remain in treatment longer.[61]

We include in patient characteristics the date of admission as well as quarterly dummies.[62] The admission date captures time trends in the data, while the quarterly dummies allow for seasonal

[61] More precisely $\frac{\overline{h}-h_o}{\mu}$ is the mean time to completion, conditional on completion. Therefore the effect of a given variable x in the expected time to completion (conditional on completion) of the average patient is given by the following expression:

$$\frac{d\left(\frac{\overline{h}-h_o}{\mu}\right)}{dx} = \frac{\frac{d\left(\overline{h}-h_o\right)}{dx}\mu - \frac{d\mu}{dx}\left(\overline{h}-h_o\right)}{\mu^2}$$

where μ and $\left(\overline{h}-h_o\right)$ take the values 0.1722 and $5.3318\ (= 5.9493 - 0.6175)$ for the average patient in the sample (see Table 8).

[62] In the estimation, all non-categorical variables are normalized to have mean zero and variance one. For example, to convert the coefficients on admission date into coefficients on date measured in *days* we need to divide by the standard deviation of the admission date variable, e.g. the time trend in the drift equation is $\frac{0.0028}{394} \approx 0.000007$ per day.

Variable Description	h_0	μ	$\overline{h}$	λ	$\ln(u)$
Date of Admission	−0.1677	0.0028	−0.2307	0.0429	−0.0458
	(0.0493)	(0.0041)	(0.0676)	(0.0242)	(0.0122)
Males	−0.1497	−0.0519	−0.1168	0.1628	−0.0134
	(0.1215)	(0.0101)	(0.1747)	(0.0558)	(0.0291)
Age	−0.0288	0.0041	0.0715	−0.1660	0.0619
	(0.0481)	(0.0036)	(0.0623)	(0.0280)	(0.0113)
White	−0.0508	−0.0336	−1.0415	−0.3077	0.0913
	(0.1255)	(0.0204)	(0.1416)	(0.1514)	(0.0796)
Married	−0.0399	0.0210	0.2055	0.0361	−0.0250
	(0.1178)	(0.0088)	(0.1619)	(0.0557)	(0.0273)
Severity=Casual	0.9098	0.0640	−0.1016	−0.6815	0.0048
	(0.1416)	(0.0181)	(0.1688)	(0.1166)	(0.0400)
Severity=Involved	0.2604	0.0353	0.1023	−0.3711	0.0777
	(0.1099)	(0.0096)	(0.1611)	(0.0609)	(0.0257)
Severity=Undetermined	0.0590	0.0381	−0.3240	−0.1622	−0.0396
	(0.1170)	(0.0124)	(0.1686)	(0.0646)	(0.0315)
Income if Stated	−0.0900	−0.0024	−0.1838	−0.1140	0.0364
	(0.0518)	(0.0040)	(0.0741)	(0.0288)	(0.0130)
Income undetermined	−0.1452	−0.0189	−0.2266	0.0755	−0.0350
	(0.1275)	(0.0107)	(0.1901)	(0.0695)	(0.0344)
OSA Primary Payer	0.4966	0.2095	1.1723	1.3428	−0.1163
	(0.1384)	(0.0184)	(0.2027)	(0.1379)	(0.0721)
Self Primary Payer	0.1516	0.2389	0.5517	1.3299	−0.2661
	(0.1302)	(0.0170)	(0.1412)	(0.1300)	(0.0689)
Medicaid or Medicare Primary Payer	0.0333	0.1721	1.2568	1.1955	0.0796
	(0.1414)	(0.0169)	(0.1773)	(0.1296)	(0.0692)
Own Private Insurance Primary Payer	−0.0374	0.2287	0.7677	1.1413	−0.0460
	(0.1558)	(0.0186)	(0.1901)	(0.1341)	(0.0684)
Infrequent problems at job at admission	−1.7527	0.0028	−1.3011	0.2117	−0.0861
	(0.1425)	(0.0151)	(0.2431)	(0.1060)	(0.0487)
Occasional problems at job at admission	−2.0174	−0.0137	−2.1575	0.5577	−0.2716
	(0.2344)	(0.0290)	(0.3101)	(0.1523)	(0.0719)
Frequent problems at job at admission	−3.0150	−0.0459	−2.7822	0.5018	−0.1858
	(0.3102)	(0.0313)	(0.3931)	(0.1316)	(0.0742)
Probation/Parole or Furloughed	1.3614	0.0405	1.6964	−0.3913	0.0884
	(0.1210)	(0.0112)	(0.2038)	(0.0617)	(0.0273)
Drunk driving	0.8835	0.1366	0.7651	−1.2203	0.2269
	(0.1116)	(0.0173)	(0.1548)	(0.0961)	(0.0272)
Concurrent Psychiatric Problems	−0.0429	−0.0505	−0.1383	0.0842	−0.0080
	(0.1356)	(0.0119)	(0.2320)	(0.0721)	(0.0422)
2nd Quarter	0.2002	−0.0136	0.1224	0.1208	−0.0315
	(0.1122)	(0.0105)	(0.1636)	(0.0666)	(0.0356)
3rd Quarter	0.2525	−0.0021	0.1058	0.0807	−0.0049
	(0.1113)	(0.0093)	(0.1614)	(0.0718)	(0.0323)
4thQuarter	0.2781	−0.0083	0.2900	0.2235	−0.1054
	(0.1133)	(0.0089)	(0.1688)	(0.0637)	(0.0303)
Unemployed (and not a student)	−0.0733	−0.0027	−0.1299	0.1580	−0.0445
	(0.0956)	(0.0085)	(0.1809)	(0.0480)	(0.0249)
Unobserved Characteristic	1	−0.0235	−0.8535	−0.9248	0.2616
	–	(0.0234)	(0.0989)	(0.1200)	(0.0486)

Table 15: Coefficient estimates for patient characteristics (X). Bootstrap standard errors in parentheses

effects. The time trend in μ is not significantly different than zero over the sample period, perhaps indicating that the improvement in the technology or an improvement in incentives to provide effective treatment was minor.[63] Over time patients became more likely to drop out, and were less likely to be abstinent at completion. For the most part, the quarterly effects are not very significant.

There is nothing particularly remarkable about the coefficients on the basic demographic variables. The sex of the patient matters, in that males, all else equal, respond worse to treatment and are more likely to dropout of treatment prematurely. Married patients do better in treatment, and older patients are less likely to dropout. Higher income patients are less likely to dropout and are less likely to be abstinent upon completion. Patients who are unemployed at admission are more likely to leave treatment prematurely.

Legal involvement matters significantly. Patients on parole or pending trial, or in treatment due to a drunk-driving offense, respond better to treatment, are more likely to be abstinent upon completion, and are less likely to drop out of treatment. These pluses could be due to the incentives associated with their legal involvement. Payment source matters in various ways. Not surprisingly, for example, patients paying out of pocket respond better to treatment than the others and receive less units of treatment.

The initial severity of the substance abuse problem impacts the treatment process in several ways. As expected, severity is a strong predictor of initial health. Dependent/dysfunctional alcohol abusers (the omitted category) do not respond as well to treatment as casual or involved users. Patients who enter treatment with problems on the job related to alcohol abuse respond to treatment

[63]For the same increase in health, the point estimate of the time trend in the drift equation implies that the average patient admitted in October 1992 would have completed treatment about 1.5 % faster if admitted a year later. OSA introduced "performance based contracting" over the sample period, using the information collected in MATS in its budget allocation decisions. It is possible that this change in information structure altered incentives. See Commons, McGuire and Riordan (1997).

σ	γ_3	γ_4	γ_5
0.7117	−0.0117	0.8010	0.8874
(0.0280)	(0.0021)	(0.2230)	(0.0803)

Table 16: Other estimated Coefficients

similarly to other patients (no problems on the job is the omitted category), but are less likely to be abstinent on completion and are more likely to exit treatment prematurely.

The last row of Table 15 contains estimated coefficients on the unobservable patient characteristic ε_i. The coefficient is normalized to 1 for the h_{i0} equation (this defines the units health is measured in). The unobserved patient characteristic is particularly important in the completion threshold, dropout, and units equations.

Other miscellaneous coefficients estimated are the variance of the Wiener process and the time dependent coefficients in the attrition model. These are reported in Table 16. Attrition, for example, is discretely higher in the first week and nearly constant thereafter, as we expected from the sample attrition histogram in Figure 3.

A.2 Estimated productivities for other patient populations

Table 17 reports productivities for every patient population in every agency using health as the output measure and controlling for both unobservable and observable characteristics of patient populations.

A.3 Abstinence as an output measure

An alternative measure of output is the expected number of discharged abstinent patients in a treated population. To compute this quantity, we determine distributions of final health status for

		AGENCY														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	All	10.96	14.95	9.33	8.36	8.45	8.71	8.09	7.55	6.27	5.67	10.17	8.17	6.61	7.59	7.87
		(0.84)	(1.35)	(0.90)	(1.05)	(1.22)	(0.89)	(0.54)	(0.54)	(0.85)	(0.75)	(1.56)	(0.94)	(0.69)	(1.05)	(1.32)
	1	12.87	16.74	9.35	9.97	10.73	9.75	8.81	8.32	7.75	7.23	10.17	9.68	7.13	8.52	7.92
		(0.90)	(1.69)	(1.29)	(1.30)	(1.87)	(1.21)	(0.78)	(0.76)	(1.03)	(0.87)	(2.07)	(1.14)	(0.86)	(1.22)	(1.83)
P	2	9.07	12.08	5.80	6.20	5.09	4.62	4.26	4.24	4.25	3.94	4.47	3.89	3.21	3.19	2.35
		(1.08)	(1.48)	(1.27)	(1.56)	(2.09)	(1.29)	(0.98)	(0.90)	(1.34)	(0.94)	(2.22)	(1.37)	(1.05)	(1.49)	(1.98)
O	3	8.03	12.16	8.38	5.75	5.10	6.41	6.23	5.86	3.88	3.30	8.87	5.30	5.35	5.63	6.72
		(1.13)	(1.38)	(0.59)	(1.05)	(1.14)	(0.92)	(0.74)	(0.70)	(0.96)	(0.89)	(1.19)	(1.32)	(0.78)	(1.17)	(0.94)
P	4	12.43	16.03	8.75	9.45	9.75	8.87	8.03	7.59	7.24	6.76	9.12	8.66	6.36	7.49	6.84
		(1.10)	(1.85)	(1.49)	(1.52)	(1.99)	(1.55)	(1.03)	(1.06)	(1.19)	(0.93)	(2.34)	(1.51)	(1.07)	(1.45)	(2.06)
U	5	13.29	17.93	11.94	10.47	10.94	11.43	10.60	9.74	8.01	7.22	13.37	10.87	8.59	10.05	10.61
		(1.36)	(1.92)	(1.29)	(1.42)	(1.45)	(1.53)	(1.00)	(0.99)	(1.15)	(1.08)	(2.19)	(1.39)	(0.95)	(1.48)	(1.78)
L	6	10.66	14.93	10.10	8.17	8.06	9.10	8.57	7.87	6.04	5.35	11.29	8.33	7.02	7.98	8.85
		(1.04)	(1.48)	(0.84)	(1.09)	(1.19)	(0.65)	(0.65)	(0.67)	(0.97)	(0.99)	(1.34)	(1.18)	(0.73)	(1.19)	(1.18)
A	7	12.29	16.34	10.14	9.66	10.21	10.20	9.34	8.67	7.47	6.82	11.38	9.92	7.56	8.98	9.01
		(0.87)	(1.51)	(1.05)	(1.12)	(1.35)	(1.03)	(0.59)	(0.63)	(0.88)	(0.76)	(1.74)	(0.99)	(0.72)	(1.07)	(1.47)
T	8	11.53	15.50	9.56	8.87	8.92	9.21	8.51	7.90	6.70	6.08	10.44	8.74	6.83	7.94	8.04
		(0.97)	(1.49)	(0.97)	(1.19)	(1.45)	(1.06)	(0.70)	(0.53)	(0.96)	(0.84)	(1.72)	(1.06)	(0.79)	(1.15)	(1.50)
I	9	14.22	17.89	9.31	11.09	12.56	10.36	9.21	8.79	8.84	8.43	10.13	10.66	7.44	9.11	7.95
		(1.23)	(2.30)	(1.71)	(1.52)	(2.33)	(1.64)	(1.16)	(1.17)	(1.14)	(0.93)	(2.58)	(1.51)	(1.17)	(1.52)	(2.29)
O	10	12.28	15.23	7.46	9.38	10.26	8.26	7.31	7.02	7.40	7.12	7.76	8.46	5.78	7.03	5.84
		(1.13)	(2.15)	(1.56)	(1.46)	(2.54)	(1.63)	(1.09)	(1.09)	(1.14)	(0.73)	(2.38)	(1.45)	(1.08)	(1.51)	(2.14)
N	11	7.96	12.18	9.22	5.83	5.29	7.04	6.90	6.35	3.98	3.35	10.30	5.85	5.95	6.33	8.15
		(1.42)	(1.74)	(0.92)	(1.23)	(1.45)	(1.38)	(1.02)	(0.98)	(1.12)	(1.10)	(1.25)	(1.60)	(0.90)	(1.35)	(1.13)
	12	13.01	17.25	10.14	10.09	10.84	10.19	9.29	8.75	7.78	7.18	11.08	9.96	7.60	8.99	8.69
		(1.10)	(1.71)	(1.26)	(1.32)	(1.55)	(1.25)	(0.81)	(0.81)	(1.04)	(0.92)	(2.10)	(0.94)	(0.85)	(1.29)	(1.77)
	13	10.71	14.24	8.36	8.00	7.65	7.71	7.13	6.67	5.91	5.40	8.57	7.13	5.62	6.35	6.28
		(0.99)	(1.54)	(1.15)	(1.35)	(1.64)	(1.24)	(0.80)	(0.80)	(1.06)	(0.90)	(1.92)	(1.13)	(0.81)	(1.24)	(1.66)
	14	12.84	16.74	9.64	10.04	10.81	10.09	9.14	8.58	7.83	7.26	10.68	10.03	7.39	8.87	8.40
		(1.04)	(1.65)	(1.20)	(1.25)	(1.70)	(1.12)	(0.74)	(0.71)	(1.07)	(0.89)	(1.97)	(1.09)	(0.82)	(1.23)	(1.70)
	15	6.99	11.04	8.15	4.87	4.15	5.63	5.61	5.28	3.14	2.60	8.66	4.38	4.98	5.04	6.69
		(1.14)	(1.60)	(0.90)	(1.03)	(1.14)	(1.10)	(0.85)	(0.88)	(0.90)	(0.87)	(1.28)	(1.35)	(0.77)	(1.22)	(0.78)

Table 17: Estimated Productivities in terms of health

each individual in the population. More precisely, for a given final health status, we compute the probability of abstinence conditional on this health status. For completors this is fairly easy, as we estimate the completion threshold, $\widehat{\overline{h}}_{ij}$. For dropouts the calculation is a bit harder because there is a probability distribution over the final health status of patients. To compute expected abstinence, we simulate from this distribution and average the abstinence probability over the simulation draws.

Table 18 reproduces the productivity comparisons of Table 11 using abstinence, instead of aggregate health improvement, as the measure of output. The productivity rankings using abstinence appear very different. In fact, the correlation coefficient between the two sets of productivities is only 0.31.[64] As the productivity numbers in this table are literally the number of abstinent patients per dollar spent, it appears that the agencies in Maine are producing between 2.5 and 5.0 abstinent

[64] The low correlation between productivities using health improvement and abstinence reported in Tables 11 and 18 may be related to the different assumptions involved regarding the counterfactuals. Assuming that untreated patient receive no health improvement may be a dramatically different assumption than assuming that they are all non-abstinent. However, the following policy analyses show that conclusions regarding good and poor performing agencies are fairly insensitive to the output measure used.

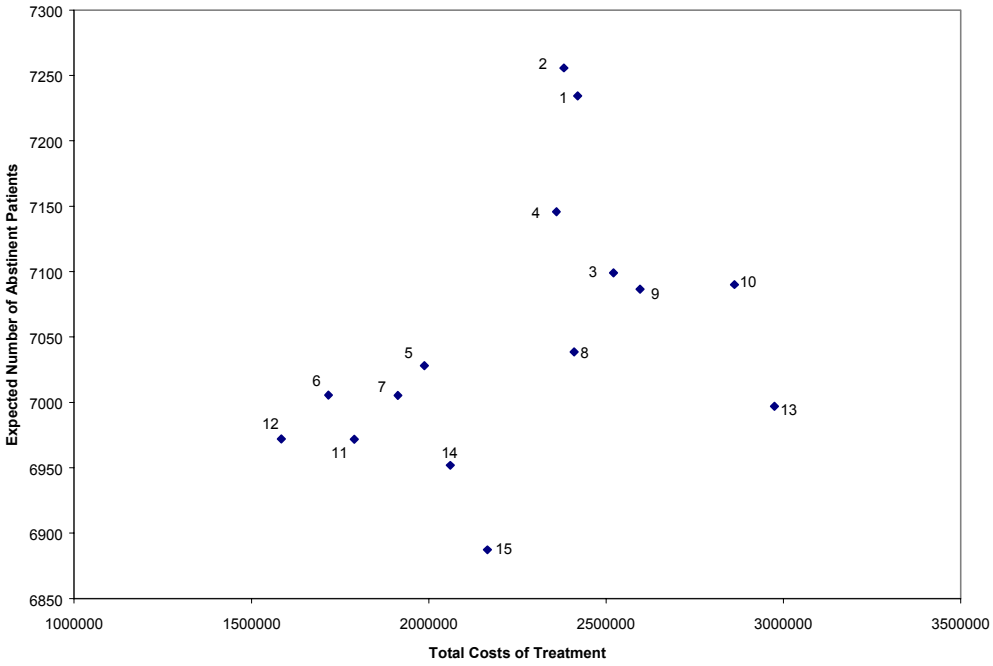

Figure 8: Expected number of abstinent patients versus costs

patients per $1000. This calculation leaves aside the question of how many of these patients would have been abstinent in the absence of any treatment.

Figure 8 repeats the analysis of Figure 7 using abstinence at discharge, rather than health improvement, as the relevant measure of output. The results are similar. In this case, the agencies that are potentially on the efficiency frontier are Agencies 2, 4, 5, 6, and 12. Agencies 13 and 10 still appear to be very inefficient.

Using bootstrap methods, we test as before whether an agency is on the efficiency frontier. Table 19 reports results for both abstinence and health improvement. Using abstinence as the output measure, Agencies 3, 8, 10, 13, and 15 can be rejected as efficient with 95% confidence, 9 and 14 can be rejected with 90% confidence. It seems clear that regardless of the outcome variable

	Own population		Entire population NOT accounting for pop. eff.		Entire population accounting for pop. eff.	
Agency	product.	ranking	product.	ranking	product.	ranking
1	0.0036 (0.0002)	8	0.0034 (0.0001)	8	0.0030 (0.0002)	10
2	0.0032 (0.0004)	9	0.0031 (0.0003)	11	0.0031 (0.0003)	8
3	0.0024 (0.0001)	14	0.0023 (0.0001)	15	0.0029 (0.0002)	12
4	0.0036 (0.0004)	7	0.0036 (0.0003)	6	0.0031 (0.0003)	9
5	0.0037 (0.0005)	5	0.0039 (0.0003)	4	0.0036 (0.0004)	5
6	0.0039 (0.0002)	2	0.0039 (0.0002)	3	0.0042 (0.0002)	2
7	0.0038 (0.0002)	3	0.0039 (0.0001)	2	0.0037 (0.0002)	4
8	0.0030 (0.0001)	12	0.0030 (0.0001)	12	0.0030 (0.0001)	11
9	0.0037 (0.0004)	6	0.0035 (0.0003)	7	0.0028 (0.0003)	13
10	0.0031 (0.0002)	11	0.0032 (0.0002)	10	0.0025 (0.0002)	14
11	0.0032 (0.0003)	10	0.0032 (0.0002)	9	0.0040 (0.0004)	3
12	0.0050 (0.0004)	1	0.0048 (0.0003)	1	0.0045 (0.0004)	1
13	0.0024 (0.0002)	15	0.0026 (0.0002)	13	0.0024 (0.0002)	15
14	0.0038 (0.0003)	4	0.0037 (0.0002)	5	0.0034 (0.0003)	6
15	0.0025 (0.0002)	13	0.0025 (0.0002)	14	0.0032 (0.0003)	7

Table 18: Agency productivity rankings using abstinence as an output measure

	Abstinence		Health improvement	
Agency	Estimate	Probability	Estimate	Probability
1	0	0.747	0	0.468
2	1	0.883	1	1.000*
3	0	0.026*	0	0.169
4	1	0.390	1	0.201
5	1	0.546	0	0.442
6	1	0.760	1	0.708
7	0	0.429	0	0.299
8	0	0.039*	0	0.039*
9	0	0.078	0	0.000*
10	0	0.013*	0	0.000*
11	0	0.383	1	0.870
12	1	0.870	1	0.740
13	0	0.000*	0	0.000*
14	0	0.0649	0	0.117
15	0	0.007*	0	0.130

Table 19: Statistical tests of being on the efficiency frontier in terms of abstinence

used, Agencies 8, 9, 10 and 13 are underperformers.

A.4 Substance abuse treatment literature

We discuss how our work relates to the literature on substance abuse treatment. The received literature focuses on four questions: Is treatment effective? Are all treatment programs equally effective? Why do programs differ in their effectiveness? Which treatments are more cost effective?

A.4.1 Is treatment effective?

This question begs for a definition of effectiveness. McLellan et al. (1997) identifies three dimensions of effectiveness: 1) reduction of alcohol or drug use; 2) improvement in personal and social functioning; and 3) improvements in public health, and conclude that the weight of the evidence supports the conclusion that treatment is effective in each dimension. Our study does not directly

address the effectiveness of treatment because our sample does not include alcohol abusers who do not receive treatment. However, our conclusion that some programs are more effective at reducing alcohol use than others (Table 6) certainly does suggest that the treatment provided by the better agencies is effective. To think that treatment is ineffective would seem to require a conclusion that the treatment provided by the other agencies is counterproductive.

Recent research views substance abuse treatment as a multi-product activity. Accordingly, McLellan et al. (1980) has proposed an Addiction Severity Index (ASI) composed of seven distinct categories of outcome measures: employment; medical status; alcohol use; drug use; legal status; family and social relationships and psychiatric symptoms.[65] Recognizing this perspective, OSA designed MATS in order to recover multiple outcomes to evaluate programs (i.e. abstinence, reduction in use, employability, job improvement, problems at job/school, problems with significant other/family, problems with the law and the judicial system, etc.).[66]

A strength of our methodology is that it utilizes the information contained in multiple outcome measures to estimate a production process and compare productivities. In addition to reduction

[65]The interdependence of these different categories of health improvement is still not well understood. For example Jaffe (1984) discusses the independence of these indicators in the short run, while Moos, Finney and Cronkite (1990) found that improvements in any one category are correlated with improvements in the other categories.

[66]MATS data is based on agency assessments of patients at admission and discharge from treatment. Some of this data are therefore self-reported measures. MATS data was collected in order to implement performance based contracting (PBC). Lu (1999) argued that agencies changed reporting practices after the introduction of PBC in fiscal year 1993. The broader substance abuse treatment literature has also questioned the reliability of self-reported data (e.g. McLellan et al (1997), Ball and Ross (1991), Long et al. (1998), Butler et. al (1987), Moos, Finney and Cronkite (1990), Maitso et. al (1990), Aitken (1986), Aplser and Harding 1991)). We would not expect this to be a problem if reporting biases in our outcome variables were not correlated with agency assignments.

The timing of measurement of our variables is also potentially problematic. MATS measures variables at admission into treatment and at discharge. Aplser and Harding (1991) criticize measures at discharge because they may not reflect the long lasting effects of treatment. On the other hand a long follow-up period has the problem of distinguishing between the effects of several treatment episodes. They also argue that measurement at admission time may be biased because of the "hit the bottom" effect, but this criticism only matters for our study if this effect varies across agencies.

in use,[67] our approach gives weight to treatment outcomes which OSA did not highlight, including time in treatment, and discharge status. We postulate an underlying latent measure of a patient's health, and view multiple outcome measures as containing different information about improvements in a patient's underlying health. In contrast to the ASI, which gives fixed weights to clinically important outcome measures, our structural model implicitly weights the importance of these measures according to their information content as revealed by the data.

Moreover, in principle we could extend our methodology to include additional outcome measures contained in our data set that we have not exploited yet. For example, our study includes "problems on the job" at admission as a patient characteristic (included in X). We could also use discharge data to include reductions in problems on the job as an outcome measure similarly to how we analyze reduction in frequency of use. Presumably this extension would increase the accuracy of our estimates or allow a richer model, although these gains would come at a cost of greater computational complexity.[68]

A.4.2 Are all treatment programs equally effective?

The relative performance of substance abuse treatment programs is controversial. Emrick (1975) argues that the literature does not support the conclusion that programs are different in their performance. However, the more recent literature repeatedly observes that substance abuse treatment programs differ in their effectiveness (McLellan et al. (1997), McLellan et al. (1993), Ball and Ross (1991), Anglin and Hser (1990) among others). For example, McLellan et al. (1993) find significant differences in the nature and effectiveness of treatments for four otherwise similar private programs,

[67] We consider frequency of use information at discharge only for patients who complete treatment. At the cost of some computational complexity, we could also consider frequency of use data for patients who leave treatment prematurely.

[68] See foonotes 10, 11 and 12.

and suggest that such differences are likely to be even more pronounced for publicly funded programs with unstable funding. Our comparison of the fifteen largest publicly funded outpatient programs in Maine confirms that the performance of these agencies differ even after controlling for observable and unobservable patient population characteristics (Table 17).[69]

A.4.3 Why do programs differ?

Some recent work in the substance abuse treatment literature tries to open the "black box" of treatment and identify its "active ingredients" (McLellan et al. (1993), McLellan et al. (1998), Ball and Ross (1991), Finney et. al (1996), Moos, Finney and Cronkite (1990)). In contrast, most of the "black box" literature relies on patient characteristics as the main predictors of differences across programs. Characteristics such as less severity of dependence, intact marriage, lower psychiatric symptoms, job, less family problems, minimal criminal activity, are associated with better outcomes (McLellan et al. (1997), Apsler and Harding (1991), Anglin and Hser (1990)). Our results conform with this literature in that we control extensively for patient characteristics. However, we go further and identify agency fixed effects as an important determinant of treatment effectiveness.

In their search for the active ingredients of treatment, researchers devoted themselves to the acquisition of very detailed data sets with information on all aspects of the treatment scenario.

[69] Our treatment of unobserved characteristics solves a sample selection problem inherent in non-randomized data, i.e. different agencies having populations with different unobservable characteristics. Sample selection problems are intrinsic to field studies, and a failure to deal with this adequately is often the basis of criticism (e.g. McClellan (1997)). In the context of our setting, Shen (1998) argued that the introduction of performance based contracting induced agencies to admit patients more selectively. Of course, even randomized trials can suffer from sample selection problems (e.g. Heckman and Smith (1995); Strohmetz, Alterman and Walter (1990)). Patient attrition creates a second selection problem in our data, we deal with it by explicitly modelling a patient attrition process. The received literature has also recognized the importance of patient attrition in evaluating outcomes (e.g. Apsler and Harding (1991), Anglin and Hser (1990), Ball and Ross (1991), Gerstein and Hardwood (1990)).

Ball and Ross (1991) exhaustive study of six methadone treatment programs, found that leadership, organization, staffing patterns, amount of services to patients were among the variables that accounted for a significant proportion of the variance across programs. Similarly, Joe, Simpson and Sells (1994) found evidence that both the type of the admission staff and the staff responsible for the treatment plan mattered for the patient relapse rates. Moos, Finney and Cronkite (1990) consider subjective indicators of treatment quality from patients and staff that measure variables such as support, relationship between staff and patients, organization of the program, etc. Our agency fixed effects can be interpreted as capturing these differences. Our conclusion that superior performance is idiosyncratic is consistent with Ball and Ross' (1991) statement regarding Methadone treatment programs: "Each clinic develops its own philosophy of treatment which appears to be strongly conditioned by the director's personal philosophy," which they found was important to explain some differences in performance.

The type and quantity of treatment services have been associated with differences in outcomes across programs (McLellan et al. (1993), McLellan et al (JAMA 1993)). For example, Moos, Finney and Cronkite (1990) use treatment intensity as an explanatory variable of outcomes in their analysis of five residential programs for alcoholics.[70] Lu and McGuire (2001), using a different sample of MATS data, found a positive and significant effect of more units of treatment for the more severe substance abusers although this effect disappears when they control for the interaction of "units of treatment" and "time in treatment."[71] Time in treatment is sometimes used as a proxy

[70] See also Finney, Hanh and Moos (1996) where they review the studies comparing outpatient and inpatient programs (which differ in intensity), and Walsh et al. (1991) interesting study based on randomized clinical trials where patients are assigned to three treatment modalities varying in treatment intensity.

[71] Other studies worth mentioning but where causality cannot be established due to endogeneity are: Moos, Finney and Cronkite (1990) who show an association between higher participation in some treatment components such as therapy sessions, AA meetings, films, attendance to church services, etc. and outcomes; and Ball and Ross (1991) who arrive to similar conclusions.

for services received under the assumption that longer stays yield a higher exposure to treatment (Apsler 1991). The literature shows that patients who stay longer in treatment tend to do better (McLellan et al. (1997), Longabaugh (1983), IOM report (1990), Tims et al. (1991), Ball and Ross (1991), Moos, Finney and Cronkite (1990).[72]) Not all of the literature recognizes that time in treatment is potentially endogenous in the sense that more difficult patients might dropout of treatment earlier (McLellan et al., 1997).[73] Our analysis endogenizes time in treatment and patient attrition within a structural model of the treatment process.

We do not have data on an individual patient's time path of treatment. However, we do know the number of units of treatment provided in each treatment episode from which we can compute the (average) intensity of treatment. A good direction for further research is to extend our structural model by adding an equation that predicts intensity of treatment.[74] We expect that treatment intensity will impact the drift and may, in part, explain some of the observed differences in productivity across agencies.

A.4.4 Which treatments are more cost effective?

Cost-effectiveness analyses of substance abuse treatment are fewer than treatment effectiveness studies. Most studies on cost-effectiveness compare alternative treatment modalities (e.g. inpatient vs. outpatient). Walsh et al. (1991), for example, find that the ultimate treatment costs for alcoholic patients receiving an initial three weeks of inpatient therapy followed by participation

[72] There is a new trend analysing the effects of "brief interventions" for alcoholics and a number of studies claim their effectiveness as compared with longer treatment. Drummond (1997) criticizes this new literature on the basis of sample selection.

[73] An exception is French et al. (1991) who control for unobserved patient heterogeneity in their study of the impact of time in treatment on labor market outcomes at a one year follow-up.

[74] The dependence of average intensity on time in treatment allows for variable actual intensity over a treatment episode. For example, patients might be more likely to miss appointments as time goes on.

in an Alcoholics Anonymous (AA) group was only 10% more than for a comparison AA group not receiving the initial inpatient treatment and performed somewhat better. Longabaugh et al. (1983) showed that a partial hospital setting was substantially less costly than an extended inpatient setting with no significant differences in performance. Long et al. (1998) compared a five week inpatient program with a two week program followed by daily outpatient therapy concluding that the shorter program reduced costs by 33% with no significant difference in outcomes six to twelve months later.[75] Our study focuses solely on outpatient services and finds significant differences in cost-effectiveness of treatment agencies. We do not explore explicitly whether these differences are due to differences in program design[76], or due to differences in efficiency (e.g selection of better clinicians.) This is a good topic for further research.

Machado (2001), using aggregate MATS data, found that agencies that spent more per patient did not have better abstinence rates at discharge,[77] suggesting that Maine could reduce expenditures on publicly funded substance abuse treatment without compromising performance. Our more detailed study of cost-effectiveness, using patient-level data, supports this suggestion and develops a methodology for benchmarking treatment agencies and identifying which are more cost effective. We conclude that Maine potentially could reduce spending on alcohol abuse treatment without compromising the health outcomes of patients by identifying and transferring best practices (Tables 1 and 12).

[75]The data came from a facility that revamped its program. The authors caution that the performance results could be due to better staff motivation in the revised program with a shorter inpatient stay.

[76]Treatment agencies provide different kinds of outpatient services. In this paper we weight "family therapy," "individual sessions" and "group sessions" according to their relative unit costs. This approach implicitly allows group therapy to be more effective than individual therapy.

[77]Machado (1998) expands her regression model to allow for heterogeneity in the use of funds across agencies and finds evidence that OSA could gain by reallocating funds to at least one agency. Her results, however, are not very powerful due to data aggregation and few observations per agency.