To assess the impact of various factors on the costs of providing care to veterans, we performed a quantitative analysis of VA patient and facility data. Our analyses were designed to incorporate the factors specified in the legislation authorizing this study, as well as other potentially important factors identified in Phase I of the study or by other researchers.

We approached our analysis with three general objectives in mind. First, we had to structure the analysis to yield clear, policy-relevant, practical recommendations for VERA. Second, the analytical approaches we used had to be methodologically sound; in particular, they had to adjust for a comprehensive set of patient- and facility-level characteristics that might influence patient care costs. Here, we believed it was important to identify whether particular variables had a statistically significant effect on the costs of care and to have the ability to explore how VISN allocations would change in response to our attempts to adjust, or control for, a wide range of variables. Third, the analysis had to be designed to accommodate the VA’s overall mission of providing health care to veterans as well as the specific objectives of VERA, particularly those related to equity and simplicity.

One of the main challenges we faced in developing our analytic approach was devising a way to incorporate patient- and facility-level factors and determining how these factors influence VISN allocations. VERA is fundamentally a capitation-based allocation system that allocates to each VISN a fixed dollar amount per patient enrolled in each of the three patient classes. Thus, we believed it was important to conduct the analysis at the patient level. At the same time, many of the factors of interest to both Congress and the VA are facility characteristics (e.g., teaching affiliations and condition of the physical plant). Consequently, we believed it was important to conduct a facility-level analysis in addition to the patient-level one.

Ultimately, however, VERA is used to allocate resources to VISNs, and any efforts made to account, or control, for additional patient or facility characteristics will affect VISN allocations. Thus, we believed our analyses should address the impact of any variable(s) on total VISN allocations. Potential modifications to “improve” VERA’s ability to explain the variation in patient costs have limited value if they increase the system’s complexity while doing little to change resource allocations at the VISN level. In essence, then, our analysis was conducted in three stages: patient, facility, and VISN.
OVERVIEW OF ANALYTIC METHODS

This section describes the motivations for our analytic approach and summarizes our methods. Subsequent sections in this chapter describe our analyses, including data and statistical methods, in detail.

Patient- and Facility-Level Equations

The first stage of our analysis focused on examining factors that affect health care costs at the patient level. Specifically, we used regression analysis techniques, described more completely below, to explain the variation in veterans’ annual costs of care—including inpatient, outpatient, and long-term care costs—as a function of sociodemographic variables, health status measures, the availability of alternative sources of care, and the facility (or facilities, in the case of some veterans) where care was delivered. Table 2.1 lists the patient-level variables.

In the second stage of the analysis, we focused on identifying treatment facility characteristics that affect patient costs. We used the estimates for the facility variables that we obtained from the patient-level regression equations as the dependent variable in a set of facility regression equations. That is, the facility-level analysis was aimed at explaining the extent to which various facility characteristics explained differences in veterans’ costs after controlling for differences in the characteristics of the

<table>
<thead>
<tr>
<th>Variable</th>
<th>Note(s) (see below)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1</td>
</tr>
<tr>
<td>Health status/case-mix measure</td>
<td>1</td>
</tr>
<tr>
<td>Income</td>
<td>3</td>
</tr>
<tr>
<td>Race/ethnicity</td>
<td>4</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
</tr>
<tr>
<td>Marital status</td>
<td>3,4</td>
</tr>
<tr>
<td>Physicians per capita</td>
<td>1</td>
</tr>
<tr>
<td>Hospital beds per capita</td>
<td>1</td>
</tr>
<tr>
<td>Rural/urban status</td>
<td>1</td>
</tr>
<tr>
<td>Distance to closest facility</td>
<td>1</td>
</tr>
<tr>
<td>Distance to closest CBOC</td>
<td>1</td>
</tr>
<tr>
<td>VA priority</td>
<td>5</td>
</tr>
<tr>
<td>Medicare reliance</td>
<td>1</td>
</tr>
<tr>
<td>Medicare imputation indicator</td>
<td>1</td>
</tr>
<tr>
<td>Medicaid generosity—general</td>
<td>3</td>
</tr>
<tr>
<td>Medicaid generosity—long-term care</td>
<td>1</td>
</tr>
<tr>
<td>Facility indicator</td>
<td>1</td>
</tr>
</tbody>
</table>

NOTES: CBOC: Community-Based Outpatient Clinic. 1Variable included in policy model; 2Variable included in base-case regression equation; 3Variable not included in policy model because it contains a potential measurement error; 4Not consistent with mission/values; 5Not consistent with current policies.
veterans that each facility serves. Thus, the facility equations attempted to explain cost differences as a function of each facility’s location, infrastructure characteristics, labor and non-labor prices, medical school affiliations, research programs, and consolidation activity. Table 2.2 lists the facility-level variables.

In the third stage of the analysis, we used the patient- and facility-level regression equations we derived in stages 1 and 2 to predict each veteran’s total annual costs, after controlling for both patient and facility characteristics. We then aggregated predicted patient costs to the VISN level to simulate how VISN allocations would vary after controlling for the variables included in the regression equations. Although we focused our analysis on VISN allocations, we could, in principle, aggregate the data in various ways. For example, we could aggregate the data by facility, state, or patient subpopulation. Here, it should be noted that the model could be used by VISN directors as the basis for making allocations to their facilities.

The three-stage structure of the analysis is attractive in that it provides a method for simulating how the VISN allocations would be affected by changes in any of the vari-

<table>
<thead>
<tr>
<th>Table 2.2</th>
<th>Explanatory Variables Used in Facility-Level Equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Note(s) (see below)</td>
</tr>
<tr>
<td>Rural/urban status</td>
<td>1</td>
</tr>
<tr>
<td>Residents per full-time physician</td>
<td>1,2</td>
</tr>
<tr>
<td>VA labor index</td>
<td>1,2</td>
</tr>
<tr>
<td>Average food cost per bed day</td>
<td>1</td>
</tr>
<tr>
<td>Energy price (dollars per million BTUs)</td>
<td>1</td>
</tr>
<tr>
<td>Contract labor costs</td>
<td>1</td>
</tr>
<tr>
<td>Square feet of building space per acre of land</td>
<td>1</td>
</tr>
<tr>
<td>Square feet of building space per unique patient</td>
<td>1</td>
</tr>
<tr>
<td>Research costs per 1,000 unique patients</td>
<td>1,2</td>
</tr>
<tr>
<td>Percentage of funded research</td>
<td>6</td>
</tr>
<tr>
<td>Average building age as of 2001</td>
<td>6</td>
</tr>
<tr>
<td>Average building condition</td>
<td>6</td>
</tr>
<tr>
<td>Leased square feet per patient</td>
<td>5</td>
</tr>
<tr>
<td>Ratio of historic to total number of buildings</td>
<td>6</td>
</tr>
<tr>
<td>Total number of buildings</td>
<td>6</td>
</tr>
<tr>
<td>Indicator for recent facility/management consolidation</td>
<td>6</td>
</tr>
<tr>
<td>Occupancy rate</td>
<td>5,6</td>
</tr>
<tr>
<td>Number of CBOCs per 1,000 unique patients</td>
<td>5</td>
</tr>
<tr>
<td>Direct patient care FTEs per 1,000 unique patients</td>
<td>5</td>
</tr>
<tr>
<td>Non-patient care FTEs per 1,000 unique patients</td>
<td>5</td>
</tr>
<tr>
<td>Long-term care beds per 1,000 unique patients</td>
<td>5</td>
</tr>
<tr>
<td>Special program beds per 1,000 unique patients</td>
<td>6</td>
</tr>
</tbody>
</table>

NOTES: Unique patient: measures the number of patients who are seen at least once at a facility during a given year (rather than the total number of visits). FTEs: full-time-equivalent employees. 1Variable included in policy model; 2Variable included in base-case regression equation; 3Variable not included in policy model because it contains a potential measurement error; 4Not consistent with mission/values; 5Not consistent with current policies; 6Not statistically significant.
ables included in the patient and facility equations. Variables can easily be added, deleted, or modified and a new set of predicted costs can be generated and aggregated at virtually any desired level.

**Fully Specified and Policy Models**

Using the patient- and facility-level regression equations, we constructed two types of models, with two distinct objectives in mind (see Table 2.3).

Our first model, which we refer to as the “fully specified model,” was intended to provide the best possible explanation of variation in patient costs. Toward this end, we tried to identify and include variables in both the patient and facility equations that we believed might influence the costs of care and for which data were reasonably available. Here, we relied on our knowledge of the relevant literature, specific concerns raised in the legislation that required the VA to undertake this study, lessons learned through the case studies that were conducted during Phase I of the project, and our prior experience modeling individuals’ health care costs.

Our second model, which we refer to as the “policy model,” was intended to be more appropriate for policy purposes than the fully specified model. In constructing this policy model, we sought to delete those variables contained in the fully specified

<table>
<thead>
<tr>
<th><strong>Table 2.3</strong></th>
<th>Descriptions of the Models Used in the Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
<td><strong>Description</strong></td>
</tr>
<tr>
<td>Fully specified model</td>
<td>Patient-level and facility-level regression equations designed to provide the best possible explanation of variation in patient and facility costs; includes variables believed to influence the costs of care and for which data were reasonably available.</td>
</tr>
<tr>
<td>Policy model</td>
<td>Patient-level and facility-level regression equations intended to be more appropriate for policy purposes than the fully specified model. Variables contained in the fully specified model were deleted if they (1) were not statistically significant; (2) were inconsistent with the VA’s mission, vision, or values; and/or (3) failed to meet current VA policy objectives. Variables related to efficiency considerations (e.g., physician FTEs per 1,000 patients) were also deleted.</td>
</tr>
<tr>
<td>Base case model</td>
<td>Regression equation-based methodology that represents an effort to take into account factors included in the current VERA allocation methodology (described in Chapter One); includes only variables that measure patient health status (three VERA patient categories), and research and education costs; and adjusts for geographic variation in labor and non-labor costs.</td>
</tr>
<tr>
<td>VERA-3 policy model</td>
<td>Policy model that uses 3 VERA patient categories for case-mix adjustment.</td>
</tr>
<tr>
<td>VERA-10 policy model</td>
<td>Policy model that uses 10 VERA patient categories for case-mix adjustment.</td>
</tr>
<tr>
<td>VERA-47 policy model</td>
<td>Policy model that uses 47 VERA patient categories for case-mix adjustment.</td>
</tr>
<tr>
<td>VA DCG policy model</td>
<td>Policy model that uses VA-modified DCG patient categories for case-mix adjustment.</td>
</tr>
</tbody>
</table>

NOTE: DCG = Diagnostic Cost Group.
model that (1) were not statistically significant; (2) were inconsistent with the VA’s mission, vision, or values; and/or (3) failed to meet current VA policy objectives. Variables related to efficiency considerations (e.g., the number of full-time-equivalent employees [FTEs] per 1,000 patients) were also deleted.

To assess the potential influence of the health status classification system on patient costs, we ran four different versions of both the fully specified and policy models, with each version relying on a different case-mix measure. These case-mix measures are described below.

**Fully Specified Model.** All widely used methods that account for interpersonal variation in health care costs adjust for at least some sociodemographic characteristics (e.g., Ash et al., 2000; Pope et al., 2000; Weiner et al., 1998; Kronick et al., 2000; Averill et al., 1999; von Korff et al., 1992). Numerous studies have shown that age and sex are correlated with health care utilization and health status (McClure, 1984; Lubitz, 1987; Hornbrook, Goodman, and Bennett, 1991; Hoff and Rosenheck, 1998; Kazis et al., 1998), in part because they are correlated with physiological developments such as degeneration of body systems. Social factors, such as marriage and education, are also predictive of health status, although the causal pathways are poorly understood (Lillard and Panis, 1996; Goldman and Smith, 2002). Thus, the fully specified model controls for age, sex, racial/ethnic origin, marital status, and income. In addition to examining the effects of sociodemographic factors on patient cost variations, we were led, by the legislation that required the VA to undertake the study, to focus on the effects of a variety of other factors. These factors included the clinical characteristics of the patients treated (i.e., case mix), infrastructure characteristics, facility and management consolidations, and facility location (urban versus rural). Figure 2.1 depicts an overview of the fully specified model.

Figure 2.1—Overview of Construction of the Fully Specified Model
“Policy” Model. In keeping with our desire to include in the policy model only variables that meet current VA policy objectives, we deleted a number of variables from the fully specified model. For example, we deleted variables related to efficiency considerations (e.g., direct patient care FTEs per 1,000 unique patients). We reasoned that statistically controlling for such variables might lead to an undesirable set of financial incentives that reward inefficient behavior. That is, if we included variables in the model related to efficiency considerations (such as the number of FTEs per 1,000 patients) over which VISN directors have some degree of control, we would risk embedding inefficiencies into the allocation scheme. Using the example of the variable for FTEs per 1,000 patients, if this variable were included in the policy model, the VA could end up rewarding VISNs that, by some measure, have too many staff members. Figure 2.2 shows the various steps that were undertaken to construct the policy model.

It is important to note that both the fully specified and policy models can potentially serve several purposes. For example, the fully specified model could be used to generate insight into the VERA supplemental, or adjustment, process. That is, because the fully specified model attempts to explain as much of the variation in costs as possible, it could be applied to assess the degree to which a VISN’s request for supplemental funding is due to factors within or beyond its director’s control.

---

**Figure 2.2—Overview of Construction of the Policy Model**
The policy model, in contrast to the fully specified model, could be used to assess the implications of various policy changes on VISN allocations. For instance, the model could be used to assess how allocations would change if VERA allocations were adjusted based on elderly veterans’ Medicare expenditures. In fact, because Medicare reliance among veterans is increasing as the veteran population ages, we have chosen to include a measure of Medicare use in our policy model, the results of which are reported in Chapter Three. Finally, in principle, the policy model could also be used as an allocation tool, replacing the current VERA system, which relies essentially on a set of national prices adjusted for differences in labor and selected non-labor costs.

Case-Mix Measures

One of the ways that VERA seeks to ensure that resources are allocated equitably is by adjusting for differences in the health status of patients within each VISN. As described in Chapter One, VERA currently adjusts for case mix by assigning patients to one of three categories—Complex Care, Basic Vested, and Basic Non-Vested—according to their level of health care use, basing capitation rates for each category on the expected costs of the care for patients in each of the three categories.

Considerable variability exists within each of the three categories with respect to actual patient costs. For example, in the Complex Care group, the national average cost for patients in the Domiciliary Patient Class 1 (i.e., patients receiving domiciliary care and meeting other criteria pertaining to long-term care utilization in the current fiscal year) in FY 2000 was $24,810, about $17,000 less than the $42,153 capitated payment for Complex Care. In contrast, the average cost for patients in the Transplant Patient Class (i.e., patients who received a heart, lung, bone marrow, kidney, or liver transplant in a VA facility within the past three fiscal years) was $78,211, about $36,000 more than the capitated payment (General Accounting Office, 2002). This variability penalizes networks that, through no fault of their own, have proportionately more transplant patients and rewards those that have proportionately more domiciliary patients.

One goal of our study was to determine whether VERA adequately adjusts for differences in case mix across the VISNs. To address this issue, we categorized patients using four different case-mix systems. We then estimated four patient-level cost models that were identical, with the exception of the case-mix system used to group patients:

VERA-3: the original three VERA categories.
VERA-10: a case-mix system with ten VERA categories.
VERA-47: a case-mix system with 47 VERA categories.
VA DCGs: VA-modified Diagnostic Cost Groups.

1Domiciliary care is a type of care that amounts, essentially, to temporary lodging for ill or disabled veterans unable to be cared for at home.
We selected the VERA-3 case-mix adjustment measure because the VA has been using it; the remaining case-mix measures were chosen because the VA has been considering using them and because the data needed to assess their potential use were readily available. In preliminary analyses, we also considered models that included Hierarchical Condition Categories (HCCs), Resource Utilization Group II scores, and Global Assessment of Functioning scores. However, we dropped these measures from our analysis because they provided little improvement in explanatory power. We did not estimate models using other case-mix adjusters, such as Adjusted Clinical Groups (Weiner et al., 1998) and Resource Utilization Group III scores (Fries et al., 1994), because the data needed to implement them were not readily available.

**VERA Patient Categories.** The VA has established a hierarchy of 47 VERA patient categories based on concepts of clinical appropriateness, policy decisions, and treatment resource requirements. Each person who enters the VA health care system is assigned to one of the 47 VERA patient categories based on the care he or she receives. Assignment to patient categories is based upon numerous data elements, including inpatient and outpatient *International Classification of Diseases, Ninth Edition, Clinical Modification* (ICD-9-CM) diagnostic codes, ICD-9-CM procedure codes, Current Procedural Terminology (CPT) codes, Resource Utilization Groups II scores, clinic stop numbers, bed specialty codes, and laboratory test results obtained from patient registries. Patients who qualify for more than one category are assigned to the highest category (in terms of expected resource use) for which they qualify.

The VA has also aggregated the 47 patient categories into 10 broader patient categories and then further combined these 10 into the three patient case-mix categories currently in use within VERA.

As we noted above, three of our patient-level equations incorporated the VERA-3, VERA-10, and VERA-47 categories, respectively, to control for the effect of case mix. Each of the three regression equations included one dichotomous variable for each VERA category except one, which served as the reference group (for example, the model that used VERA-10 included dichotomous variables for nine of the ten patient categories.) Appendix B shows how these three classification systems compare.

**VA DCGs.** Our fourth patient-level model used a modification of the DCG system to control for case mix. The HCC/DCGs system, a widely used case-mix adjustment approach developed by researchers at Boston University, uses approximately 15,000 ICD-9-CM diagnosis codes to classify patients into 545 clinically homogeneous groups called DxGroups. The DxGroups are further collapsed into clinically homogeneous condition categories (CCs) that require similar resources. To avoid double counting within a disease category, a hierarchy is established based on disease severity, so that patients are assigned only to the highest ranked CC (the HCC) among sets of related conditions. For example, a woman diagnosed with both breast cancer and secondary metastasis to the lung would be assigned to the metastatic cancer HCC but not to the breast cancer HCC. The number of times the ICD-9-CM

---

2Since the VA updates the list periodically, the number of categories at any given point in time may be slightly higher or lower than 47.
code is recorded throughout the year and the site of treatment (inpatient versus outpatient) do not affect assignment (Ash et al., 2000; Pope et al., 2000).

The VA has modified HCC/DCGs to reflect differences between the veteran population and the privately insured population, for which off-the-shelf HCC/DCG software is intended. Specifically, the VA combined 30 HCCs (those that are very uncommon in the VA population or do not predict significant positive costs) into one category and added 14 VERA category flags for special disability programs (e.g., spinal cord injury, traumatic brain injury, and serious mental illness). The VA then predicted the costs for each patient from the HCC model and assigned patients to one of 24 “VA DCG” categories based on their predicted costs (VHA Executive Decision Memo, 2001). In our DCG patient-level equation, one dichotomous variable was included for each VA DCG except the highest-cost VA DCG, which served as the reference group.

Predicting Patients’ Costs

As indicated above, the patient and facility equations were used to generate a predicted annual cost for each VA patient in a given year. These predicted costs represent our best estimate of what each patient’s costs would be after controlling for the independent, or explanatory, variables included in the patient and facility regression equations. Thus, predicted costs were determined by the set of variables included in the patient- and facility-level regression equations. We generated the predicted patient costs for each of the four case-mix measures described above and aggregated these costs to the VISN level.

DATA SOURCES

Our analyses relied on individual-, facility-, and county-level data. The individual-level data set was prepared by VHA’s Allocation Resource Center (ARC), using a set of specifications supplied to the VA by RAND staff. The ARC data set contains information on the annual costs of treating patients at each VA facility, along with a host of socioeconomic, eligibility, cost, and health status variables. The cost measure included in the data set was taken from the VA’s “Cost Distribution Report” and is based on individuals’ VA health care use. The ARC assigns costs to patients according to the methodology described in its “Patient Costing Manual” (VA Allocation Resource Center, 2000).

In addition, the patient-level file contains information on individual Medicare “reliance,” state-level Medicaid generosity, and county-level health care resources. Patient-level data on annual Medicare expenditures for users of the VA health system (which we refer to as “Medicare reliance”) were available to RAND through agreements with the Centers for Medicare & Medicaid Services (CMS) and the VA’s Management Science Group (MSG).

Individual-level data on Medicaid expenditures for patients in the VA health care system are not readily available. As a substitute, we created state-level measures of Medicaid generosity using data on state-level Medicaid expenditures and the number of poor adults in each state. The data on Medicaid expenditures came from the
Health Care Financing Administration (HCFA)-2082 reports (now collected as part of the Medicaid Statistical Information System, or MSIS). The data on the number of poor adults in a state were taken from the Kaiser Family Foundation web site (www.kff.org), which are based on estimates from the Census Bureau’s Current Population Survey.

Information on the local (i.e., county-level) supply of physicians and hospital beds was obtained from the Area Resource File (ARF). The ARF data are produced annually by Quality Resource Systems, Inc., under contract to the Health Resources and Services Administration.

The majority of the data for the facility-level equations came from either the ARC or VA headquarters. Again, these files were constructed based on a set of specifications that RAND submitted to the VA. The facility-level file contains data on each facility’s structural characteristics, costs, and staffing levels. We supplemented these data with information on state-level energy prices from the Department of Energy’s State Energy Price Report and information on the rural or urban status of the location of the parent VA facility from the ARF.

We requested and received both individual and facility data for FY 1998 through 2001. However, because HCCs and VA DCGs were not available for 2001, we limited our analysis to data from the earlier fiscal years. Moreover, after testing the stability of the estimates over time, we chose to limit our analysis further to a single year of data. As such, the results presented in Chapter Three of this report are based on data from FY 2000, the most recent year for which data were available on all patient-level variables.

Because some veterans received care at more than one facility, we aggregated each veteran’s costs across facilities to obtain one observation per person. Each facility in which the veteran received care in that year was assigned a weight in the analysis equal to the proportion of the total cost of the care they received at that facility.

**DEPENDENT AND EXPLANATORY VARIABLES**

In this section, we describe the dependent and explanatory variables that we used in the patient- and facility-level regression equations. As indicated previously, the same dependent variables were used in the fully specified and policy models. However, the fully specified model contained a larger set of explanatory variables than did the policy model.

**Dependent Variables**

The main focus of our analysis was on explaining how patient and facility characteristics affect the costs of providing health care to veterans. Consequently, the dependent variable used in our patient-level cost equation was the VA’s annual cost of providing health care to the individual. The costs included all medical care costs (inpatient, outpatient, and long-term care), as well as education and research support costs, resident salaries, equipment costs, and NRM costs.
The dependent variable for the facility-level equations came from the independent variables in the patient-level cost equations that served as indicators for the 143 major facilities—generally, acute care hospitals—in the VA health care system.\textsuperscript{3,4} By including these facility indicator variables in the patient-level regression equation, we were able to estimate the effect of being treated in a particular institution on each patient’s costs. These facility-specific cost shifts provide information on the variation in costs across facilities if characteristics of the patients treated are held constant. These estimated cost shifts served as the dependent variable in the facility-level regression equations. More precisely, the facility-level equation was used to understand the degree to which the cost shifts attributable to facility differences can be explained systematically by the facility characteristics detailed below.

We note one issue that may be important for interpretation. To estimate the cost of medical care at the patient level, ARC begins with the total dollars per budget unit (e.g., total dollars in a Cost Distribution Report account) and then divides that amount by patient utilization within that account (e.g., number of inpatient days) to arrive at an estimate of the cost per unit of utilization (e.g., cost per day of hospitalization) (VA Allocation Resource Center, 2000). Each patient’s utilization is then tabulated and assigned costs so that patient-level costs can be estimated. This method ensures that the dollar cost of the relevant VA unit equals the ARC-estimated costs of utilization within that unit. However, the budget of a VA unit is not necessarily identical to the economic costs of producing the medical care products and services that were used by VA patients within that unit. As a result, our dependent variable can be thought of most appropriately as being derived from relative value weights for the underlying health care used by VA patients, rather than as estimates of the absolute economic cost of production.

**Explanatory Variables**

Table 2.1 (previously shown) lists the explanatory variables used in both the patient- and facility-level regression equations. The table also shows which variables are included in the fully specified and policy models. At the outset, we included 17 categories of variables in the fully specified patient-level equation and 22 categories in the fully specified facility-level equation. After we deleted variables that failed one or more of the criteria for inclusion in the policy model, 11 variable categories remained in the patient-level regression model, and 9 remained in facility-level regressions.

It is important to note, as indicated previously, that we assessed the potential effects of various case-mix measures across facilities by running the fully specified and policy models using four different case-mix measures. However, these case-mix mea-

\textsuperscript{3}Our count of the number of VA facilities reflects a wave of recent management consolidations. The VA health care system contains more than 143 major physical structures, but these structures are organized into only 143 management units. In addition to the 143 facilities, the VA has over 450 outpatient clinics. However, each clinic reports through a major facility that is ultimately responsible for how each patient’s care is managed.

\textsuperscript{4}Technically, indicator variables for only 142 facilities were included in the equation because one facility served as the reference category.
sures were not included in the facility-level regression equations; rather, they were included in the individual-level equations, which, in turn, produced the facility-specific cost shifts that the facility equations attempt to explain.

**Description of Selected Variables in the Regression Equations**

Many of the variables included in the patient- and facility-level regressions are straightforward and do not warrant discussion (e.g., age, race/ethnicity, sex). However, some of the variables require more-detailed descriptions.

**Patient-Level Equation.** The patient-level equation contains two measures of county-level health care resources: hospital beds and physicians per capita. These measures were taken from the ARF and were matched to individual veterans based on their home zip code. Similarly, the measures of distance to the facility at which the individual was treated and to the closest CBOC were calculated using the home zip code of the individual and the zip code of the facility or CBOC. We used these variables to explore whether the availability of other health care resources in the county in which the veteran resides and the distance the veteran must travel to receive VHA services affect the amount of care the veteran obtains from VHA facilities.

Medicare reliance was measured as the percentage of total health care costs (Medicare payments, including beneficiary cost-sharing amounts, plus VA costs) that is covered by Medicare. A person is said to be more reliant on Medicare as this percentage increases. In addition, in the fully specified model, the patient-level equation includes two measures of state Medicaid generosity. To obtain measures of generosity that are relevant for the VA population, we first created a general measure that is based on state-level Medicaid expenditures on recipients who are eligible for coverage because they are elderly, blind, or disabled. To incorporate information about a state’s breadth of coverage, we scaled the expenditures by the number of poor adults (age 18 and over) in the state. The resulting measure (expenditures per poor adult) incorporates both aspects of program generosity: spending and eligibility. The second measure of Medicaid generosity has the same basic characteristics but focuses specifically on long-term care. In this case, the measure was calculated as state-level Medicaid expenditures on long-term care for each poor elderly adult (age 65 and over).

The facility variables in the patient-level regression indicate the facilities where an individual was treated. The facility indicators take values between zero and one and measure the percentage of an individual’s total annual VA costs that were incurred at each facility. Approximately 80 percent of veterans in the data set were seen at only one facility during the year. For these individuals, the indicator variable for the facility where they received treatment had a value of one, and all other facility indicators were zero. Similarly, veterans who were treated at multiple locations had multiple facility variables (the number coinciding with the number of facilities at which they were treated).

---

5 The distance is calculated from the center of the home zip code to the center of the facility’s zip code. The precise methodology used to calculate these distances came from Meridian World Data and is described on its web site (www.meridianworlddata.com/HTML9/distance-calculate-2.asp).
were treated), each taking values greater than zero and less than one. Since the value of the facility variables for an individual is based on the share of total costs incurred at each facility, the sum across all these variables for an individual will equal one.\footnote{The facility variables incorporated are very similar to the prorated patient (or PRP) calculations that the VA currently uses in its methodology for counting workload.}

**Facility-Level Equation.** In the facility-level regression equations, the VA labor index is a VISN-level variable generated by the VA to measure the difference in wages across geographic areas. The VA labor index is used to adjust allocations in the current system.

The facility-level equation also includes a measure of the average physical condition of the buildings at the facility. It is measured on a scale of 1 to 5, with higher scores indicating better physical condition. These data were taken from the VA’s “Capital Asset Baseline Assessment.”

Also included in the facility-level equation are several variables aimed at measuring medical education and research activity related to academic affiliations, based on the findings from our review of the literature (see Phase I report, Wasserman et al., 2001). Education costs can be measured in a number of ways. To look at direct costs, one could use a categorical variable for the number of residents per facility to test whether costs might vary by residency program size. To assess the impact of teaching on the provision of patient care services by teaching physicians, one could construct a variable based on the ratio of residents to physicians per facility. This variable would measure the intensity of physician involvement in teaching activities (that is, the higher the resident to physician ratio (or the more residents per physician), the more involved physicians are in teaching activities) and would account for the net impact of residents on physician productivity. While teaching activities reduce the time teaching physicians can devote to patient care activities, residents also provide patient care. When we tested these two different measures of medical education programs, we found that the intensity measure had more explanatory power. Thus, we selected that variable for inclusion in the regression equations. Finally, we constructed two variables to measure research intensity. One measured total research costs per 1,000 unique patients; the other was expressed as the percentage of funded research that took place at each facility.

**DATA CLEANING AND IMPUTATION**

In this section, we describe the steps we took to clean and prepare the data for analysis.

**Individual Data**

In general, the data that were obtained for the patient-level analysis were complete, clean, and deemed reliable. However, for some variables, missing data were a problem. When possible, we used information from other years or other observations on
the same person within the same year to logically impute values for the missing variables. This method was used in cases where the variable value for an individual would be unlikely to change over time (e.g., sex and race) or would change in a predictable fashion (e.g., age). In cases where we were unable to logically impute a value for the variable, it was coded as missing. In the patient-level regressions, all variables were entered into the equation categorically rather than as continuous variables. This methodology allows the missing values to be coded as such and to be included in the analysis.

Data on individual-level Medicare expenditures were needed to generate the measure of Medicare reliance used in the patient-level equation. Unfortunately, expenditure information is not available for individuals who are enrolled in Medicare managed-care HMO (Medicare+Choice) plans. However, for such individuals, the data do indicate the number of months the individual was enrolled. We used this information, as well as information on the VA facility (or facilities) where the person was treated, to impute a Medicare expenditure for people in Medicare+Choice plans. A timing issue also arises in the calculation of Medicare reliance. The most recent Medicare expenditure data that were available to us were from FY 1999. As such, the FY 1999 data on Medicare expenditures (including the imputed values for the HMO enrollees) were brought forward to use in our cost equations for FY 2000. We inflated the FY 1999 Medicare expenditures into 2000 dollars using the Medical Care Consumer Price Index (CPI) to make them comparable with the VA cost data. In addition, we imputed Medicare expenditures for those individuals who are Medicare eligible (age 65 and over), but for whom we have no Medicare expenditure data in FY 1999. This group consisted primarily of individuals who became Medicare eligible during FY 2000. Because we did not have any information on the number of months the newly eligible were enrolled, we imputed six months of Medicare costs for this group.

---

7For example, if information on sex was missing for an individual in the FY 2000 data, we looked at data for FY 1998, FY 1999, and FY 2001 to see if sex was reported in another year. If data for another year had information on the individual’s sex, then we assigned that information to the FY 2000 observation. This sort of logical imputation is particularly useful for variables such as sex that we would not expect to change over time.

8For example, instead of entering an individual’s age into the model directly, we generated a set of indicator variables representing different age groups (e.g., less than 25 years of age, 25–34, 35–44, etc.). These indicator variables take the value of 1 if the individual falls into that age category and 0 otherwise. Each person will fall into only one age category. When entering the variables in this fashion, we can include an additional category for those individuals for whom the age variable is missing.

9The imputation procedure assigns the facility-specific average fee-for-service Medicare expenditure to individuals treated at that VA facility and who were enrolled in Medicare HMOs. The average is scaled to reflect the number of months the individual was enrolled in the HMO.

10In doing so, we have implicitly assumed that Medicare reliance is stable over time, at least in the short run.

11For these imputations, facility-specific average expenditures (based on fee-for-service clients) were assigned to individuals for whom no FY 1999 Medicare information exists.
Facility Data

The data that were obtained to estimate the facility-level regression equations came from a wide array of sources within the VA. While individual data elements were relatively clean and complete, combining the data elements was difficult. The problem stems primarily from recent management consolidations, which led to facility information being reported at different levels in different systems. For example, much of the information on facility infrastructure was measured in FY 2001 and reflected the consolidation of facilities. However, the data on costs and staffing patterns still include information for individual facilities that have since been consolidated.12 Moreover, the patient-level data show people being treated in these facilities. As such, we spent a great deal of time going through the facility data and aggregating them (or disaggregating them in some cases) to generate data for a consistent set of facilities. In cases where facilities were missing some data elements, we assigned the median value for that variable across all facilities.

STATISTICAL TECHNIQUES

The overall goal of the analyses was to evaluate the potential impact on health care costs of various patient- and facility-level characteristics. To this end, we implemented a two-step regression procedure. The first step in our analysis was to estimate a patient-level multivariate equation of annual VA health care costs:

\[ C_i = X_i \beta_1 + H_i \beta_2 + A_i \beta_3 + L_i \beta_4 + W_i \theta + \epsilon_i \]  

(1)

where

- \( C_i \) is equal to the total annual costs for veteran \( i \);
- \( X_i \) is a vector of sociodemographic variables for veteran \( i \);
- \( H_i \) is a vector of health status (or case mix) variables for veteran \( i \);
- \( A_i \) is a vector of availability of local health care resources (that is, physicians and hospital beds) for veteran \( i \);
- \( L_i \) is a vector of geographic location variables for veteran \( i \);
- \( W_i \) is a vector of VA facility variables indicating the percentage of veteran \( i \)'s annual costs that were incurred at each facility;
- \( \epsilon_i \) is an error term that is independently and identically distributed (i.i.d.) \((0, s^2)\);

and \( \beta_1 - \beta_4 \) and \( \theta \) are parameters to be estimated.

In general, annual health care costs have a very skewed distribution (i.e., a large number of patients with low costs and a long right tail representing a small number

---

12In most cases, the consolidation reflects only a change in management organization and does not indicate that the consolidated facility has been closed.
of patients with very high costs), and the VA data are no exception. When this skewness occurs, standard linear regression using the dollar value of annual costs as the dependent variable may not fit the data very well. As an alternative, researchers often use alternative specifications such as ordinary least squares (OLS) regression with a log or square root transformation of the dependent variable or gamma regression with a log link (Manning and Mullahy, 2001).

For the present analyses, however, we have chosen to use standard, OLS regression models, in the interests of simplicity, transparency, and consistency. One of the goals of the VERA system is for the allocation framework to be simple and predictable. We have developed our analytic strategy with this goal in mind. Models with log or square root transformations, and gamma regressions, are more complicated than the linear model, and the results from these specifications are more difficult to interpret directly. Moreover, the linear specification is used in much of the existing literature on risk adjustment. This consistency with the previous literature is desirable because it provides the context for the interpretation of our results.

The parameter estimates associated with the facility indicator variables in the patient-level regression equations reflected the cost associated with being treated at a particular facility after controlling for all observable individual characteristics. These estimated facility cost shifts set up the second step in our empirical analysis. In this step, we sought to explain the variation in these facility cost shifts using facility-level characteristics. In these regression equations, the estimated coefficient of a particular characteristic measures its association, on average, with a shift in annual costs. This equation has the following general form:

$$\hat{\boldsymbol{\theta}}_j = L_j'\delta_1 + S_j'\delta_2 + P_j'\delta_3 + I_j'\delta_4 + M_j'\delta_5 + R_j'\delta_6 + E_j'\delta_7 + v_j$$

where

- \(\hat{\boldsymbol{\theta}}_j\) are the estimated facility-specific average shifts in annual costs, based on the results of the patient-level equation;
- \(L_j\) is a vector of geographic location-related measures for facility \(j\);
- \(S_j\) is a vector of medical school affiliation measures for facility \(j\);
- \(P_j\) is a vector of labor and non-labor prices or costs for facility \(j\);
- \(I_j\) is a vector of measures of the physical plant/infrastructure for facility \(j\);
Data Sources and Methods

$M_j$ is a vector of consolidation-related measures for facility $j$;

$R_j$ is a vector of research- and education-related measures for facility $j$;

$E_j$ is a vector of efficiency-related measures for facility $j$;

$\nu_j$ is an error term that is i.i.d.(0,$s^2$); and

$\delta_1 - \delta_7$ are parameters to be estimated.

We also estimated the facility-level equations using OLS regression.

The first and second stages of the estimation were linked. The facility cost shifts that serve as the dependent variable in the second stage equation were estimated in the patient-level equation. As a result, a separate set of facility cost shifts was estimated for each specification of the patient-level regression, and thus, a separate set of facility-level results is generated. Therefore, changes in the specification of individual-level equations could lead to changes in the estimates of the impact of facility-level characteristics on the facility cost shifts, even if the specification of the facility-level equation does not change. For example, the estimated impact of academic affiliations in the second stage could vary depending on the specific set of health status measures used in the first stage.

The estimates from the regression equations identify the factors, both patient- and facility-level, that have a significant impact on costs. However, these estimates do not directly address the question of how VISN allocations would be changed if these variables were considered in VERA. To address this question, we used the regression estimates to simulate VERA allocations to VISNs under various scenarios.

The first step in the simulation process was to generate predicted annual costs at the patient level. To do this, we worked backwards through the estimation process and started with the facility-level regression, where the estimated coefficients were used to predict facility cost shifts. The predicted cost shifts were then used in combination with the patient-level parameter coefficient estimates to generate predicted annual costs for each individual veteran. The empirical implementation of this process is illustrated in Equations 3 and 4. Patient-level annual predicted costs are given by the following

$$\hat{C}_i = X_i^\prime \hat{\beta}_1 + H_i^\prime \hat{\beta}_2 + A_i^\prime \hat{\beta}_3 + L_i^\prime \hat{\beta}_4 + W_i^\prime \bar{\theta}$$ (3)

where the symbol “^” represents an estimated coefficient and $\bar{\theta}$ is a vector of predicted facility cost shifts determined by the following equation.

$$\bar{\theta}_j = L_j^\prime \hat{\delta}_1 + S_j^\prime \hat{\delta}_2 + P_j^\prime \hat{\delta}_3 + I_j^\prime \hat{\delta}_4 + M_j^\prime \hat{\delta}_5 + R_j^\prime \hat{\delta}_6 + E_j^\prime \hat{\delta}_7$$ (4)

The predictions are based only on the variables included in the equation. Thus, unobserved factors such as efficiency and quality of care are not directly taken into account. The same is true for predictions based on the policy model, where some variables are excluded based on the criteria that were outlined previously. In this case,
the predicted costs generated from the policy model directly take into account only those factors that are consistent with the VA’s mission and current policy.\textsuperscript{14}

The second step in the simulation process involved aggregating the predicted patient-level costs to the VISN level. Because some veterans incur costs in multiple VISNs during the year, we broke out each individual’s annual predicted cost across facilities based on the share of actual costs that were incurred by that individual at each facility. This allowed predicted costs to be aggregated to the facility and VISN levels. The VISN-level aggregate can be interpreted as an estimate of the costs the VISN would be expected to incur based on the characteristics of the individuals the VISN treats and the characteristics of the facilities.

The VISN-level predicted costs could then be used to generate a simulated allocation for any lump-sum appropriation. To do this, we used the VISN-level predicted costs to calculate the proportion of total predicted costs incurred by each VISN. The share estimates can then be applied to any given appropriation to derive the associated VISN-level VERA allocations. The calculation of the simulated allocation for a particular VISN is illustrated in Equation 5.

\[
\text{Allocation for VISN}_i = \frac{\text{Predicted Cost for VISN}_i}{\sum_{i=1}^{22} \text{Predicted Cost for VISN}_i} \times (\text{Appropriation})
\]

To interpret the simulation results, it is useful to have a basis of comparison, or benchmark allocation, against which each simulation can be judged. The actual FY 2002 allocation is perhaps the most obvious benchmark. However, the comparison between the simulated allocations and the actual FY 2002 allocations confounds two different effects: (1) the difference in methodology (regression versus workload counts and national prices) for determining the allocations and (2) the difference in the patient- and facility-level characteristics that are included in the models. In an effort to separate out these two effects, we made three different comparisons. First, we developed a “base case” regression equation that includes only variables that are currently considered in the VERA methodology: the three VERA patient categories, the labor index, research costs, and education costs. The results from these regressions were then used to simulate a base case allocation for each VISN. Because the base case model includes only variables that are used in the current VERA system, the comparison between the base case and actual FY 2002 allocations isolates the difference that is due to the methodology used.

The second type of comparison we made was between the simulated base case allocations and the simulated VERA-3 policy model allocations. These comparisons

\textsuperscript{14}However, if variables that are omitted from the regression are correlated with both the dependent variable and one or more of the included covariates, the estimated coefficients for the included covariates will be biased. In other words, the coefficient estimates will pick up some of the effects of the excluded variables. As such, the policy model may indirectly take into account some characteristics not included in the model.
show the impact of the additional patient and facility variables that we controlled for in our regression equations, using the same case-mix variables.

The third set of comparisons was designed to isolate the effect of the alternative case-mix measures on VISN-allocations. Here we compared the VERA-3 policy model simulations to those based on the VERA-10, VERA-47, and VA DCG case-mix specifications. Since the case-mix measure that is used represents the only difference among these models, the results of these comparisons will illustrate the effect of the alternative case-mix measures on VISN allocations.

SENSITIVITY ANALYSES

In addition to using the four case-mix measures described above, we conducted a number of analyses to determine how sensitive our findings were to alternative data and model specifications.

First, while our analysis relied largely on patient-level data from the year 2000, we also estimated the patient-level equations using data from 1999.

Second, we used alternative measures of each patient’s total annual treatment costs as the dependent variable in the patient-level regression equations. As indicated previously, the person-level cost data used in the primary analysis came from the ARC. However, the VA has recently decided to rely on its Decision Support System (DSS) database for budgetary and management purposes. Because significant differences exist in the ways in which ARC and DSS estimate patients’ costs, we tested whether our regression and simulation results would be sensitive to these cost allocation differences. We further tested the sensitivity of our results to changes in the ways in which costs are allocated to patients by using patient-level cost data provided by the VA’s Health Economics Resource Center (HERC). In conducting the

---

15ARC estimates patient-level annual costs from data in the Cost Distribution Report (CDR), Outpatient Clinic File (OPC), Patient Treatment File (PTF), and other sources. A limitation of the ARC methodology is that it does not estimate the cost of individual health care encounters. Per diem inpatient costs, for example, are estimated based on the total dollars per CDR account divided by the total bed days of care for each bed service (VA Allocation Resource Center, 2000); the total per diem costs for a particular veteran over the course of a fiscal year are estimated by summing up the per diem costs from different bed services. As Barnett (under review) notes, this approach assumes "that the cost of medical surgical days is proportional to the length of stay; information on the relative cost associated with [each] DRG [Diagnosis Related Group] is not employed."

ARC allocates trainee, education, research support, and other indirect costs to patients “on a per diem basis for inpatient-related accounts and on a cost per Clinic Stop basis for outpatient-related accounts” (VA Allocation Resource Center, 2000).

16In contrast to ARC, HERC estimates the cost of each health care encounter, then sums them up to obtain an estimate of the cost of each patient in each fiscal year. As summarized by Barnett (under review),

The cost of acute medical and surgical care was estimated using measures of relative value estimated from a cost-function created from Veterans’ stays in Medicare hospitals (Wagner et al., 2002 [under review]). The cost of long-term care was based on estimates of the relative resource use associated with case-mix measures from periodic assessment of VA long-term care patients (Yu et al., 2002). The cost of outpatient visits was estimated using the payments from Medicare and other payers as a measure of relative value (Phibbs et al., 2002 [under review]).
cost allocation sensitivity analysis, we created a subset of the patient-level data set that included only those patients who received care at a single VA facility during FY 2000, which amounted to roughly 80 percent of the patients. Limiting the sample in this way provided a reasonable test of the sensitivity of our results, while circumventing the time-consuming effort that would have been required to account properly for patients treated at multiple facilities under the DSS and HERC cost allocation methods.

Finally, to test the sensitivity of our results to the veteran population used in the analysis, we estimated the patient- and facility-level equations after including all Priority 7 patients. We initially excluded the Basic Care Priority 7 patients from our analysis, because they are excluded from VERA workload calculations. However, we were interested in assessing the extent to which the patient and facility equations’ parameter estimates and the simulated VISN allocations changed if these patients were included in the analysis.

Like HERC, DSS estimates the cost of individual health care encounters. However, the DSS methodology is quite different from HERC’s. The HERC website (www.herc.research.med.va.gov, 2002) summarizes the DSS methodology as follows:

DSS extracts costs from the VA payroll and general ledger. These are assigned to departments based on periodic reports made by managers, who assign costs of the six categories of expense to departments. Some sites use time reports and accounting data instead of managerial reports to assign costs to departments. The calculation of department costs from the managerial estimates, payroll, and general ledger data is done by the DSS program called the Account Level Budgeter (ALB). Overhead (the cost of departments that do not produce patient care) is distributed to patient care departments using a step-down method. Direct cost or the number of square feet of occupied space are used as the basis of the distribution. Costs of intermediate products are then determined. Examples of intermediate products are a chest x-ray, a unit of blood, a 15-minute clinic visit, or a day of stay in the intensive care unit. They are called intermediate products to distinguish them from the final product, a patient encounter, which is a bundle of intermediate products.

Both DSS and HERC “normalize” cost estimates to the VA’s cost allocation system. That is, costs are multiplied by a constant factor so that, when aggregated, the dollar costs across patients sum to the relevant VA budget allocation. Because the VA budget allocation is not necessarily identical to the aggregate economic cost of producing the medical care products and services that were used by VA patients, the DSS and HERC estimates, like the ARC cost estimate, should be thought of as being derived from relative value weights for the underlying health care used by VA beneficiaries, rather than as estimates of the absolute economic cost of production.