THE SENSITIVITY AND SPECIFICITY OF FORECASTING HIGH-COST USERS OF MEDICAL CARE

By

Richard Thomas Meenan

A THESIS

Presented to the Department of Public Health and Preventive Medicine

and the Oregon Health Sciences University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Public Health

February 1999

School of Medicine
Oregon Health Sciences University

CERTIFICATE OF APPROVAL

This is to certify that the M.P.H. thesis of

Richard Thomas Meenan

has been approved

[redacted]

[redacted]D.

Merwyn R. Greenlick, Ph.D.

[redacted]

Donald K. Freeborn, Ph.D.

[redacted]

Mark Helfand, M.D., M.P.H.

Associate Dean for Graduate Studies

TABLE OF CONTENTS

LIST OF FIGURES

Acknowledgments

I want to thank my friend and colleague Maureen C. O'Keeffe-Rosetti, M.S., for her invaluable analytical and programming support during the course of this research project. I also want to thank my thesis committee (Mark C. Hornbrook, Ph.D., Merwyn R. Greenlick, Ph.D., Donald K. Freeborn, Ph.D., Mark Helfand, M.D., M.P.H.) and colleagues, Donald J. Bachman, M.S., Michael J. Goodman, Ph.D., Paul A. Fishman, Ph.D., and Karen Seccombe, Ph.D. for their helpful scientific and editorial comments.

This research was supported by Cooperative Agreement #17-C-90864/9-01, "Reforming Medicare Risk Payment through Competitive Market Forces," from the Health Care Financing Administration (HCFA), by Grant #033039 from the Robert Wood Johnson Foundation, and by a grant from Kaiser Permanente. I received additional support from the Kaiser Permanente Center for Health Research and by an Agency for Health Care Policy and Research Post-Doctoral Fellowship (#T32 HS00069). The statements contained in this thesis are solely mine and do not necessarily reflect the view or policies of HCFA, Kaiser Permanente, HealthPartners, or Group Health Cooperative of Puget Sound. I assume full responsibility for the accuracy and completeness of the information contained in this thesis.

Abstract

OBJECTIVE. The study objective is to evaluate 3 risk-assessment models as forecasters of individuals at relatively higher risk of incurring large health care expenditures. The models are the Global Risk-Assessment Model (GRAM) developed at the Kaiser Permanente Center for Health Research, a logistic version of GRAM, and a prior-expense model. GRAM was originally developed using administrative data from 3 managed care organizations (MCOs) for use in adjusting Medicare payments to health plans for enrollee health risk.

An underlying assumption of this study is that early identification of "high-cost" individuals improves the efficiency of care by encouraging prompt intervention at lower cost with similar or superior effectiveness. It is hypothesized that risk-assessment models will be effective preliminary screens for high-cost enrollees within MCO populations by exploiting the administrative datasets of such organizations.

METHODS. The sample of 98,985 individual-level cases was drawn from random samples of memberships of 3 staff/group HMOs. Demographic and diagnostic risk factor data were from 1992 and expenses were measured for 1993. Each risk-assessment model produced a distribution of individual-level annual expense predictions for comparison to actual values. (The logistic model generated predicted probabilities of high-cost status.) Prespecified high-cost thresholds (e.g., top 5%, 10%, and 25% of actual annual expense) were set within the actual and predicted distributions to analyze each model's ability to distinguish high- and low-cost status (rather than forecast individual-level dollar expense). Prediction stability was analyzed through bootstrapping methods.

RESULTS. GRAM discriminates better than its comparators between high- and low-cost enrollees over the range of possible thresholds (although the models are similar for policy-relevant thresholds). All models predict the highest-cost cases relatively well. GRAM is the most sensitive within chronic and serious disease groups that appear amenable to early intervention, although its comparators perform better on other criteria (e.g., specificity, total accuracy) within these groups.

CONCLUSIONS. This study demonstrates the utility of risk-assessment models as potential screens for future high-cost enrollees within managed care populations. However, analyses of large multi-year datasets are needed to explore the full potential of these models. As risk-assessment models become more sophisticated, their utility for high-cost case identification will be enhanced.

Introduction

Continuing pressure to restrain health care costs, intensified by spreading capitated risk contracts, has generated an interest in efficiency throughout the U.S. medical care sector. One manifestation is the demand for disease and care management (D/CM) programs especially by managed care organizations (MCOs). D/CM is intended to provide selected patients with continuous monitoring of illness episodes, access to timely and appropriate medical care, and cost-effective disease prevention and screening. D/CM may potentially contribute to efficiency by: (1) interrupting the disease course early when more cost-effective treatments are available; (2) improving continuity of care to reduce complications; (3) minimizing fragmentation and duplication of care; and (4) substituting less expensive inputs for the same process of care. Although the actual contribution of D/CM to efficiency and improved outcomes is subject to debate (Maryland HCUI 1996, Boyd et al. 1996, Warren et al. 1996, Fitzgerald et al. 1994, Ferguson and Weinberger 1998, Nash 1998), methods of screening D/CM candidates remain relevant research topics because identification early in the disease course is critical to the realization of potential program efficiencies (Mukamel et al. 1997, Brody et al. 1997).

By exploiting information contained in automated administrative and clinical datasets, regression-based risk-assessment models represent a potentially efficient screening methodology. This study compares the ability of 3 such models to prospectively forecast individuals and groups at risk of incurring large medical care expenses within an MCO's general population. Comparator models are (1) the Global Risk-Assessment Model (GRAM) developed primarily at the Kaiser Permanente (KP)

Center for Health Research (CHR) (Hornbrook et al., unpublished manuscript); (2) a logistic version of GRAM; and (3) a prior-expense model. Models such as GRAM were originally developed to adjust payments to MCOs for differences in enrollee health risks.

Specifically, we evaluate the ability of these models to distinguish between prespecified groups of "high-cost" and "low-cost" enrollees based on their actual annual medical care expense. The "diagnostic test" is a forecast of high-cost or low-cost status based on annual expense predicted by a given model. Our results outline the potential of risk-assessment models such as GRAM to act as efficient adjuncts to more detailed but resource-intensive screening techniques like social surveys or medical records review. Note that our objective is not to predict individual dollar expense, but to accurately forecast an individual's status as low- or high-cost using existing risk-assessment instruments designed originally for risk adjustment. As risk-adjusted payment systems become more common, plans should understand the potential internal applications of risk-assessment models.

Background

Since the 1960s, the medical care sector of the U.S. economy has grown substantially in absolute and relative terms. Further, the distribution of these increasing expenses has been highly skewed—the most costly 10 percent of U.S. citizens accounted for 70 percent of total 1996 medical care expenses (Getzen 1997). Subpopulations (e.g., Medicare beneficiaries, Maryland Medicaid) exhibit similar distributions (Kaiser Medicare Policy Project 1997, Maryland HCUI 1996). D/CM program growth is largely a response to these skewed expense patterns and their implications for efficiency. However, some authors claim that "traditional" D/CM has not approached its potential

efficiency because of segmentation between inpatient, insurance-based (i.e., service-limited), and community- or home care-based D/CM (Qudah and Brannon 1996, Phillips-Harris 1996). This taxonomy implies fragmentation and duplication, with no single provider responsible for delivering all needed services.

To reduce the inefficiencies of care segmentation, MCOs have begun development of integrated or population-based D/CM programs, which adopt the premise that medical care should be managed across all settings to promote wellness, prevent illness, and coordinate care. Proponents of integrated care argue that acute (inpatient) care is usually episodic, reactive to patient needs, and diagnosis-driven. This contrasts with the holistic approach of integration, which involves continuous monitoring and anticipation of patient needs (Qudah and Brannon 1996). Examples of integrated programs are the Social Health Maintenance Organization and the Program of All-Inclusive Care of the Elderly (Abrahams 1990, Leutz et al. 1994, Shen and Iversen 1992).

An important element of realizing whatever efficiency benefits from integrated D/CM exist is the efficient identification of "high-risk" patients. Measures of "risk" vary across D/CM program types and target populations (e.g., sentinel event clusters to identify patients at risk of re-admission (Roblin et al., unpublished manuscript), a frailty measure to identify nursing-home eligible patients (Brody et al. 1997), and annual per capita utilization rates and expenses to identify resource-intensive cases (Levkoff et al. 1992)). Although D/CM programs are also accountable for improving health outcomes and access, cost reduction is the most commonly promoted objective (First Health 1997, Henderson et al. 1988, LoBianco et al. 1996, Private Healthcare Systems 1996). One public-sector example is the Maryland Medicaid High Cost User Initiative (HCUI),

which tried to reduce "unnecessary" expenses while improving quality of care by substituting less costly home- and community-based services for hospital and institutional care, reducing hospital readmissions, coordinating somatic and mental health care, and arranging support services that promote treatment compliance (Maryland HCUI 1996).

Every D/CM program must identify the cases of interest, e.g., those patients most likely to generate the largest expenses. The simplest (and usually least costly) method of identifying program candidates is to select all those in the population with a particular condition such as cardiovascular disease or diabetes, which, on average, generates substantial medical care costs for the typical MCO or purchaser. Self-reported health surveys are one way to locate these individuals, although this can be expensive. Administrative data systems are also potential principal or complementary information sources (Pronk et al. 1997, Robinson et al. 1997). Here, case identification is typically pursued as part of a plan's population-based health promotion initiatives. Yet, including *all* diabetics or heart patients in an intensive D/CM program could easily exhaust its resources depending upon disease prevalence in the relevant population (Henderson et al. 1988).

The level of potential health or financial benefits from D/CM differs across patients. For example, among chronically ill patients, age influences utilization and expenses, although psychosocial adjustment to one's condition also matters (Watt et al. 1997). Freeborn et al. (1990) found similar results among elderly HMO members. Consistently high medical care users were generally older and sicker than consistently low users. They also reported more medical conditions and greater psychological distress. Such heterogeneity induces programs to segregate their populations by "risk"

level and target more intensive interventions toward higher risk levels (Qudah and Brannon 1996).

In theory, expense risk may not correlate well with other important risks. For example, expense risk and mortality risk may rise together at low absolute levels of mortality risk, but as mortality risk increases, expense risk may decline because fewer beneficial interventions exist. However, as suggested above, efficiency concerns lead most programs to target expense risk either explicitly or implicitly through the relation of utilization to expense. Further, they often use the previous outcome level, e.g., high prior expense or utilization, for identification. Prior utilization can predict future utilization well, but the opportunity to lower costs or improve health status is reduced since the adverse outcome has in a sense already occurred.

To maximize the efficiency of prompt intervention, individuals at risk should be identified, when possible, *before* large expenses arise. Yet, medical care expenses have varying degrees of predictability. Accident-related trauma expenses are largely unpredictable, especially for individuals. Acute illnesses can be costly, but are often transitory and do not persist long enough to predict future utilization. Although even trauma victims benefit from improved medical intervention, treatment options are limited for such largely unpredictable events and therefore meaningful cost reductions may focus on primary prevention, e.g., safer cars. Hence, identifying these conditions is of limited value because they are unlikely to be predictable soon enough to implement cost-effective alternative treatments.

Chronic diseases that permanently change or reduce body system functioning seem more likely to produce predictable expense patterns amenable to D/CM (Hornbrook

et al. 1985). Indeed, medical expense risk derives from the underlying *permanent* propensity to need and use health care services in the next year (Hornbrook and Goodman 1991). A basic premise of risk models like GRAM is that only permanent health status indicators, e.g., chronic disease or use propensity (including behavioral risk habits), that are essentially uncontrollable by the individual or plan are appropriate for risk assessment. Their stability, consistency over time, and relation to utilization and expense make such indicators useful for case identification (and D/CM) as well. (Distinctions between illness types should be interpreted in context. Knee or shoulder injuries, for example, may recur randomly or may suggest higher medical risk if they occur to an athlete.)

This suggests that prospective identification systems capable of recognizing characteristics of predictable expense are preferred to contemporaneous or retrospective systems, particularly if system administrative costs are limited. Examples of prospective systems include HMO mail surveys to identify the frail elderly (Brody et al. 1997), the use of administrative data to identify sentinel event clusters suggesting treatment-related re-admission risk (Roblin et al. 1997), and brief questionnaires that identify high risk of repeated hospital admissions (Boult et al. 1993) or total annual expenditures per plan enrollee (Mukamel et al. 1997). Regression-based risk-assessment models can also serve as a prospective identification system. Typically, such models have been used to design systems for health plan payment adjustment based on enrollee medical risk, e.g., Ambulatory Care Groups (ACGs – Weiner et al. 1996), Diagnostic Cost Groups (DCGs-Ellis et al. 1996), and GRAM.

A risk-assessment model tries to "explain" health status-related variation in annual per capita health care expense. It forecasts individual annual expense, but the true test of precision in the risk adjustment context is the accuracy of its expense predictions across groups (e.g., health plans, Medicare beneficiaries). In a sense, risk-assessment modeling for risk adjustment tries to equate the per capita actual and predicted expense of plans or enrollee groups. Payments are expected to better reflect the plans' "true" per capita expense risk, limiting the effects of adverse selection. Indeed, group expense is highlighted in the definition of medical risk of Hornbrook and Goodman (1991): the *expected value* of the distribution of per capita costs of efficiently-provided preventive, diagnostic, and therapeutic health care services delivered to a defined *group* of enrollees for a specific future period.

These authors point out that the ability of risk-assessment models such as GRAM to predict the expected value of an expense distribution is related to the modelers' understanding of the overall shape of the distribution. In particular, non-random group selection can lead to MCOs with relatively cheaper or more expensive populations. Hence, useful risk-assessment models should be able to identify factors underlying the skewed expense distribution. We want to exploit this ability in identifying the cases in the high-cost tail. Yet, we should note that the policy objective of risk assessment as used in risk adjustment is to modify the size of insurance payments to health plans. (Indeed, note that efficient health care is assumed in the medical risk definition of Hornbrook and Goodman 1991.) In our context, the policy objective of risk assessment is not to equate actual and predicted plan expense, but to *lower* average plan expense by facilitating the development of a more efficient "production function" through D/CM.

Material and Methods

Global Risk-Assessment Model (GRAM)

The basic form of GRAM is an additive weighted least squares regression (Neter et al. 1996):

Annual health plan expense$_{t,i}$ = f (Demographic and diagnostic risk factors$_{t-1,i}$, Error$_{t,i}$)

where t = year and i = 1,2, ... n individuals.

In this analysis, 1992 risk factor data were used to predict 1993 expense.

Data and Disease Classification

Our sample was the complete data set of 98,985 individual-level observations that was used in the initial estimation of GRAM. These data were drawn from three large non-profit HMOs in the upper Midwest and Pacific Northwest. The study population for GRAM was defined as all subscriber unit members eligible in October 1992 and March 1993 and who were covered by an outpatient pharmacy benefit. Spouses and dependents had to be eligible at least one month each year, not necessarily under the same subscriber. Thus, persons with partial data were included. Random subscriber unit samples from the participating HMOs yielded the final usable sample of 98,985 individuals. All enrollment groups, including Medicare, Medicaid, employer group, COBRA, and individuals, were included; hence, the data represent an entire HMO membership from newborns to the frail elderly. Expense is plan-specific as collected from existing automated data systems.

Primary diagnostic data were collected from each medical encounter, including ambulatory visits, urgent care, emergency, and inpatient stays. Diagnoses in GRAM are

organized within an updated and expanded KP Clinical-Behavioral Disease Classification System (C-B), which maps ICD-9-CM diagnoses into three major divisions, 19 diagnostic classes, and 117 subclasses according to "typical" clinical attributes and providers' likely behavioral responses, rather than actual care patterns (Hurtado and Greenlick 1971, DHHS 1980). Figure 1 illustrates the C-B class structure. Within C-B categories, the Clinical Resource Intensity (CRI) system groups diagnoses with similar expected resource intensity, which we assume is positively correlated with expense. Physician raters assigned ICD-9-CM codes within each C-B subclass to "resource intensity" categories for months 0-12 and months 13-24 following first diagnosis:

Extremely High (> $10,000)

Very High ($5,000-$10,000)

High ($2,000-$5,000)

Possibly High (> $2,000; no specific upper limit)

At Most Average (≤ $2,000)

The raters did not predict incurred expense. These dollar thresholds were intended to standardize expected differences in cost perceptions across specialties, e.g., we do not expect "extremely high" expense to imply the same concept to a neurosurgeon and a dermatologist. Most diseases were rated as relatively more costly during months 0-12 because that is when costs typically cluster. Although hierarchies were created within C-B groups (i.e., a person could only be in the most expensive CRI class within a C-B group), that person could be in multiple C-B groups if he or she received multiple diagnoses during the risk-assessment year (here, 1992).

Characteristics of GRAM

Since about 10% of the sample had only partial-year eligibility through voluntary termination or death, we annualized the dependent variable in GRAM by:

Actual expense * (12 / eligible months) (Ash et al. 1989)

We downweighted annualized partial-year expense cases to compensate for their relatively greater variance (Goodman et al., unpublished manuscript). Full-year cases (90% of the sample) received a weight of one. Persons with partial-year eligibility were weighted by $w(x) = (\text{months of eligibility}/12)^X$. Estimated coefficients represented risk coefficients for a population with the observed mix of full and partial eligibility among elderly (age > 64) and non-elderly groups. (Clusters of "non-use" accompany most voluntary terminations, while high-use clusters accompany most deaths, common among the elderly.) The chosen elderly and non-elderly weights were $X = 1$ and $X = 1.8$, respectively. Although introducing these new parameter values increased the risk of overfitting, they also minimized the Medicare and non-Medicare prediction errors produced by GRAM, which are important validation criteria in GRAM's original context of risk-adjusting payments to health plans.

C-B/CRI groups are represented by dichotomous dummy variables, which are turned "on" if a category has one or more coded diagnoses. Note that the final 14 demographic and 83 (of 350 possible) C-B/CRI variables in this version of GRAM are the result of an extensive model development process, which cannot be fully described here. However, it involved excluding certain C-B categories (e.g., Symptoms) on policy grounds; establishing CRI hierarchies within C-B subclasses; and combining CRI categories to produce stable C-B/CRI regression coefficients that represent the annual per

capita incremental cost of a disease in a particular C-B/CRI group with other conditions and demographics held constant. This interpretation of regression coefficients suggests that serious and/or chronic diseases should be associated with relatively higher costs than other diseases. Yet, for D/CM purposes this does not address the problem of the potentially large program expenses that may result from targeting everyone in a particular disease group, regardless of severity or utilization. Of greater interest to program implementers is the model's discriminative ability, i.e., its ability to distinguish between high- and low-cost cases as defined *a priori*. With this information, they can then search for homogeneous disease groups within the sets of correct high- and low-cost predictions.

For each observation, the predicted value of the dependent variable ($Expense_{t,i}$) represents a "score", i.e., the annual per capita health care expense expected in year *t* based on an individual's year *t-1* risk factors. An individual score is simply the sum of the relevant coefficients for that individual including the intercept. For example, GRAM predicts that a 55-year-old Medicare-disabled female receiving 1992 diagnoses of non-serious skin cancer and a sprain would generate 1993 expenses of $5,941 = $682 (female aged 50-62) + $1,777 (Medicare-disabled) + $2,846 (non-serious skin cancer) + $401 (sprain) + $235 (intercept).

Description of "Diagnostic Test"

The actual and predicted expense distributions are each rank-ordered from high to low. Then, a policy-relevant threshold distinguishing high-cost and low-cost individuals is set within each distribution, e.g., the most expensive 5, 10, and 25% of cases. We call the proportion of "true" high-cost cases correctly forecasted the model's sensitivity and the analogous proportion of true low-cost cases its specificity. The proportion of high-

cost predictions that are correct is the model's positive predictive value (PPV) and the proportion of low-cost predictions that are correct is its negative predictive value (NPV). We expect that a decisionmaker would establish the desired thresholds within each distribution based on budgetary constraints, organizational values, and attitudes about the risk of false positives and negatives.

Prior-expense and Logistic Specifications

The prior-expense model regressed non-annualized 1993 expense on 1992 expense. The logistic version of GRAM converted actual expense to high-cost status (=1) and low-cost status (=0) before estimation. Note that to produce a "level playing field" across models, our comparisons include logistic results for threshold probabilities that generated an absolute number of high-cost predictions equivalent to that generated by GRAM and prior-expense.

Results

Descriptive Statistics

Table 1 lists descriptive statistics for the sample of 98,985 individuals. In 1993, 14.2% of the sample generated no actual health care expense. Percentile values indicate the high skewness of the annualized expense distribution. In the actual sample 99% of expense values are below $19,750, but the most costly 1% range from $19,750 to $277,000 (25% of total sample expense). Predicted expenses are similarly skewed.

Forecasting Results

Figures 2-4 are receiver operating characteristic curves that illustrate the tradeoff between true positives (sensitivity) and false positives (1 – specificity) across different thresholds for various high-cost prevalence rates. The upper left corner of the figures

reflects perfect accuracy, i.e., 100% true positives and zero false positives. The figures suggest that GRAM discriminates better overall between high- and low-cost cases than its comparators. However, these figures illustrate the discrimination of each model across virtually the entire spectrum of thresholds (e.g., 2% high-cost to 98% high-cost). For more likely policy-relevant thresholds (e.g., 5% high-cost), which are near the lower left corner of the figures, the forecasting performance of the models is quite similar.

Since expenses differ between individuals, we would also like an indication of the models' ability to predict across the spectrum of high-cost cases. One summary statistic is the proportion of total "high-cost" dollars represented by correct "high-cost" predictions. Table 2 illustrates that for each model, correctly predicted individuals represent a higher proportion of total "high-cost" dollars than total high-cost cases (i.e., sensitivity), although the performance of the prior-expense model is somewhat inferior in this regard. This suggests that the impact of false negatives is probably mitigated somewhat because correctly predicted high-cost cases tend to be the highest-cost cases. A D/CM program manager interested in locating the highest-cost cases may choose to set the actual threshold high and the prediction rule low, which would mean that the program will include a larger proportion of non-targeted cases as the "price" of identifying the targeted cases.

It is also important to note that GRAM generally underpredicts the actual expense of high-cost individuals and overpredicts for low-cost individuals. The range of predicted values is narrower than the range of actual values because model risk factors do not account for random or idiosyncratic expense variation. In other words, the predicted expense score represents the *average* expected health care cost of someone with given

demographics and diagnoses. However, the model need not predict individual expense accurately if it can identify relative differences in medical risk, i.e., it produces similar orderings of actual and predicted expense.

Bootstrapping Results

A risk-assessment model should predict reasonably well across different populations. Applying a prediction model to the same database used to estimate it biases its accuracy upward because the model will capitalize on all variance associated with the predictor variables including measurement error and random individual sample-specific variation. Splitting the sample into estimation and validation groups is a common technique; however, certain high-cost conditions in our sample had too few cases to either generate a stable coefficient or assure their appearance in both data groups. Hence, we used all cases for estimation, leaving no available cases for validation.

Instead, we used the multi-sampling technique of bootstrapping (Efron and Gong 1983). From the estimation data set of 98,985 observations, we created 100 resamples of 98,985 observations each by random selection from the original sample with replacement. This procedure implies that some cases appear in no sample, while others appear in multiple samples or multiple times in the same sample. The result is a series of resamples with compositions and expense distributions that differ from the original estimation dataset and each other. We then re-estimated the model 100 times and calculated means and standard deviations of the distribution of sensitivity and specificity values across the 100 runs. Table 3 presents the point estimates and the 90 percent bootstrap confidence intervals for sensitivity and specificity generated by GRAM, using the percentile method (Mooney and Duval 1993). The point estimates of GRAM's

sensitivity and specificity appear quite stable, as do plan-specific point estimates (not shown).

Results by Disease Grouping

Tables 4-6 list prediction statistics for the 15 C-B/CRI categories in each model with the highest within-group sensitivities (5% thresholds). Most categories are identical across all models. (Orderings by PPV are qualitatively very similar.) Within these categories, GRAM's sensitivity is somewhat higher although the comparators generally do better on other criteria. (Of course, it should be noted that the prior-expense model does not actually include C-B/CRI categories; the results reported in Table 5 are simply an *a posteriori* evaluation.)

The 15 most "sensitive" C-B/CRI groups from the version of GRAM used in this analysis are included within the 25 most expensive groups as indicated by coefficient magnitude and contain each of the 10 most expensive groups. Excluding obvious conditions such as continuing pregnancies, the majority of categories among these 15 most sensitive disease groups are chronic and/or serious conditions that appear qualitatively to be both resource-intensive and amenable to D/CM or other early intervention programs. This information can prompt further analysis (e.g., comorbidities, functional health status surveys) within the group of interest. Table 7 lists these 15 most sensitive C-B/CRI groups and representative diseases within each. Note that manifestations of diabetes, a specific condition targeted for D/CM by the participating HMOs, appear in 2 of these categories.

Discussion

Many characteristics of risk-assessment models such as GRAM that make them useful for risk adjustment make them conceptually appropriate as forecasters of high-cost status as well: their foundation in a logical and coherent disease classification scheme (e.g., the C-B/CRI system); their discriminative ability among diseases; the capture of the full medical risk spectrum; and a focus on HMO technology. These characteristics should enhance the models' overall predictive power as well as their ability to identify relatively homogeneous disease groups in terms of expense risk. Our empirical results clearly support the use of risk-assessment models for this purpose, particularly in conjunction with population-based D/CM programs. GRAM shows superior overall discrimination to either a logistic or prior-expense model (although they are more similar for likely policy-relevant thresholds) and predicts the highest-cost cases relatively well. However, within chronic and/or serious disease groups, GRAM appears to sacrifice overall accuracy to maximize the sensitivity of its results.

The performance of the models we tested broadly compares to that of Brody et al. (1997) who used a self-report screening instrument (and limited administrative data) to predict elderly HMO members at risk of frailty in the next year. Although their methodology is significantly different (e.g., a dichotomous dependent variable, frailty), they report the sensitivity, specificity, and predictive values of their models. In a population with 14.6 percent frail members, they report sensitivity of 51 percent, specificity of 98 percent, PPV of 79 percent, NPV of 92 percent, and total accuracy of 91 percent for a logistic model with four variables. The predictive performance of this model was only slightly less than a similar model with 13 variables. We do not want to

overstate the comparability of their results with ours, however, because these authors focused on 5,800 elderly Social HMO enrollees in KP's Northwest Division.

Our tests used individual predicted scores; however, we do not feel that risk-assessment models such as GRAM are appropriate for directly identifying individuals for specific clinical interventions. They are more suited to give decisionmakers preliminary information about high-cost disease groups that will support further analysis. MCOs can use GRAM, for example, as a primary (or complementary) method of population needs assessment. GRAM would allow examination of retrospective administrative data for indicators of conditions associated with subsequent high cost that could be targeted within the current population. Large MCOs (e.g., enrollee populations above 100,000) might choose to re-estimate the risk-assessment model on their own data, generating predicted expense scores based on MCO-specific model coefficients. Stable plan-specific results suggest that smaller organizations with more limited resources can simply use the existing GRAM coefficients, which are based on a complete HMO population, to "score" their own population for testing. Criteria supporting more detailed investigation will depend on the decisionmaker's organizational values, risk attitudes, and budgetary constraints. The MCO might look for disease groups with good separation between high and low costs, i.e., relatively few false negatives and false positives.

Plans with a limited D/CM budget may want to focus on screening thresholds that maximize PPV to assure that identified cases are "truly" high-cost (recognizing that some deserving cases will be missed). Plans with larger D/CM budgets may choose instead to focus on sensitivity to assure identification of the largest number of high-cost cases.

Given available resources, false positives can then be reduced through more detailed screens as discussed below.

It should be noted that the current state of risk-assessment models as used in risk adjustment somewhat limits their usefulness for case identification. Note, for example, that the one-year prediction horizon limited GRAM's prediction performance because of the timing of acute events (including treatment-related complications (Mark 1996)). Because timing of first diagnosis is not available in these data, cases with quite different actual expenses may produce similar predictions, i.e., similar acute events may appear differentially predictable. Because most acute diseases are ultimately transitory, models that capture chronic conditions that permanently alter functioning and that generate persistent use levels across years are usually preferred for forecasting in the context of risk adjustment.

Yet, a prior-expense model that ignores underlying risk factors also relies on persistent use for prediction. We assert that efficiency is most greatly enhanced by early identification while expenses are relatively low. This suggests that persistent use itself may be less important in this context than the reliability of low-cost "early warning signs" of future expense. A potential screening advantage of models like GRAM over prior-expense models in particular is that they may be able to detect these signs, especially if multiple years of data are available.

Further, "outlier" expenses tend to regress toward their group mean (Goodman et al. 1991). In a prior-expense model, the population is the group; therefore, regression to the mean can have potentially large effects on the temporal stability of predictions. (By the "temporal stability" of a forecast, we mean that a forecast of "high-cost" status made

over one year would match a forecast made over multiple years.) A risk-assessment model that produces relatively homogeneous subgroups mitigates this effect. We would expect predictions to be temporally stable if the subgroup mean were clearly in one expense-risk class or the other. For screening purposes, we feel that regression to the mean should be of limited importance unless the subgroup variance is relatively large and/or expenses regress rapidly to the subgroup mean.

Also, medical and expense hierarchies in GRAM that produce stable coefficients dampen the effect of multiple low-cost conditions that may cumulatively produce a high-cost case. Hence, we probably missed some of these cases in our analysis. Further, GRAM's additive structure incorporates global effects of different primary comorbidities, which enhances its predictive power. However, it does not currently include secondary diagnoses that would allow, for example, expense differences between diabetic and non-diabetic cardiac patients to be estimated (unless patients were diagnosed separately during the risk-assessment year with diabetes and heart disease). Future explorations of GRAM and other risk models should include analyses of secondary diagnoses.

In addition, GRAM's results were based on nearly 100,000 observations, but these were still not enough to measure the relative risk of certain rare conditions. Data sets with millions of observations are needed to evaluate such conditions accurately. Further, illness must be validated in GRAM by legitimate ICD-9-CM codes and must appear as the primary reason for an encounter or stay. Hence, GRAM and other risk models are subject to the general limitations of ICD-9-CM codes and plan-specific labeling and coding practices (Iezzoni 1997).

Also, because risk models such as GRAM do not include unmet needs, they cannot provide information on persons who do not interact with the health care system and thus limit true population-based case identification. However, this offers an opportunity to supplement their results with other techniques, especially for MCOs with larger D/CM budgets. For example, surveys focused on functional and psychological domains of health could be administered to individuals within groups that the risk model previously identified as high-cost (Brody et al. 1997). Within this framework, GRAM's global population focus and use of administrative data make it an effective preliminary screen for intervention candidates requiring more detailed, resource-intensive identification.

Finally, most risk models currently omit direct measures of behavioral risk factors, functional health status, and use propensity. Even the best models do not capture behavioral risk patterns before they register as diagnosed conditions. As risk adjustment-based research explores such enhancements, the additional explanatory power should also improve the models' case identification ability.

Summary and Conclusions

We believe that this analysis clearly indicates the potential of risk-assessment models for forecasting high expense-risk within MCO-based populations. Hence, risk models such as GRAM can be extremely useful additions to the stock of available tools for locating those patients at the greatest risk of incurring large health care expenses. As risk-assessment models become increasingly sophisticated in the risk adjustment context, their utility for the identification of cases that are suitable for disease and/or care management programs will also be enhanced.

References

Abrahams R. The social/HMO: case management in an integrated acute and long-term care system. CARING Mag 1990;41:30-40.

Ash A, F Porell, L Gruenberg, E Sawitz, A Beiser. Adjusting Medicare capitation payments using prior hospitalization data. Health Care Fin Rev 1989;10:17-29.

Boult C, B Dowd, D McCaffrey, L Boult, R Hernandez, H Krulewitch. Screening elders for risk of hospital admission. J Am Geriatric Society 1993;41:811-817.

Boyd ML, B Fisher, AW Davidson, CA Neilsen. Community-based care management for chronically ill older adults. Nurs Mgmt 1996;15:155-184.

Brody KK, RE Johnson, LD Ried. Evaluation of a self-report screening instrument to predict frailty outcomes in aging populations. Gerontologist 1997;37(2):182-191.

Efron B, G Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Statistician 1983;37:36-48.

Ellis RP, GC Pope, LI Iezzoni, JZ Ayanian, DW Bates, H Burstin, AS Ash. Diagnosis-based risk-adjustment for Medicare capitation payments. Health Care Fin Rev 1996;17:101-128.

Ferguson JA, M Weinberger. Care management programs in primary care. J Gen Intern Med 1998;13:123-126.

First Health, Special care case management. 1997 Oct 22. Available from: URL: http://www.firsthealth.com/sccm.html

Fitzgerald JF, DM Smith, DK Martin, JA Freedman, BP Katz. A case manager intervention to reduce readmissions. Arch Intern Med 1994;154:1721-1729.

Freeborn DK, CR Pope, JP Mullooly, BH McFarland. Consistently high users of medical care among the elderly. Med Care 1990;28(6):527-540.

Getzen TE. Health economics: fundamentals and flow of funds. New York: John Wiley & Sons; 1991.

Goodman MJ, DW Roblin, MC Hornbrook, JP Mullooly. Persistence of health care expense in an insured working population. In: Hornbrook MC, editor. Advances in health economics and health services research, vol. 12. Greenwich, CT: JAI Press Inc; 1991, p. 149-173.

Goodman MJ, PA Fishman, MC Hornbrook, TP Gilmer, AF Nelson. Termination-related partial eligibility: implications for risk model development. Unpublished manuscript, 1997.

Henderson MG, BA Souder, A Bergman, AF Collard. Private sector initiatives in case management. Health Care Fin Rev 1988;9(Supp.):89-95.

Hornbrook MC, AV Hurtado, RE Johnson. Health care episodes: definition, measurement and use. Med Care Rev 1985;42:163-218.

Hornbrook MC, MJ Goodman. Health plan case mix: definition, measurement, and use. In: Hornbrook MC, editor. Advances in health economics and health services research, vol. 12. Greenwich, CT: JAI Press Inc; 1991, p. 109-146.

Hornbrook MC, RT Meenan, DJ Bachman, MC O'Keeffe-Rosetti, MJ Goodman, PA Fishman, AV Hurtado, MR Greenlick, AF Nelson. Forecasting health plan liability for defined populations: a global risk assessment model. Unpublished manuscript, 1998.

Hurtado AV, MR Greenlick. A disease classification system for analysis of medical care organization, with a note on symptom classification. Health Serv Res 1971;6:235-250.

Iezzoni LI. Assessing quality using administrative data. Ann Int Med 1997;127(8(2)):666-674.

The international classification of diseases, 9th revision, clinical modification. Washington, DC: U.S. Department of Health and Human Services; 1980.

Kaiser Medicare Policy Project, Medicare chart book. Menlo Park, CA: The Henry J. Kaiser Family Foundation; 1997.

Leutz WN, MR Greenlick, JA Capitman. Integrating acute and long-term care. Health Aff (Millwood) 1994;13(4):58-74.

Levkoff SE, T Wetle, E Coakley, J Kasper. Determinants of expenses and use of services among elderly HMO enrollees. HMO Pract 1992;6(3):6-12.

LoBianco MS, ME Mills, HW Moore. A model for case management of high cost Medicaid users. Nurs Econ 1996;14(5):303-314.

Mark DB. Implications of cost in treatment selection for patients with coronary heart disease. Ann Thorac Surg 1996;61:S12-15.

Maryland Medicaid High Cost User Initiative, Care management and cost savings annual report FY 1996. Baltimore: Center for Health Program Development and Management—UMBC; 1996.

Mooney CZ, RD Duval. Bootstrapping: a nonparametric approach to statistical inference. Sage University Paper series on quantitative applications in the social sciences, series no. 07-095. Newbury Park, CA: Sage; 1993.

Mukamel DB, CC Chou, JG Zimmer, BM Rothenberg. The effect of accurate patient screening on the cost-effectiveness of care management programs. Gerontologist 1997;37(6):777-784.

Nash DB. A clarion call for case managers. J Gen Intern Med 1998;13:144-145.

Neter J, MH Kutner, CJ Nachtsheim, W Wasserman. Applied linear regression models. 3rd ed. Chicago: Richard D. Irwin, Inc.; 1996.

Phillips-Harris C. The integration of primary care and case management in chronic disease. Qual Manage Health Care 1996;5(1):1-6.

Private Healthcare Systems, PHCS utilization management. 1996 Nov 2. Available from: URL: http://www.phcs.com/products/um.html

Pronk NP, P O'Connor, G Isham, C Hawkins. Building a patient registry for implementation of health promotion initiatives: targeting high-risk individuals. HMO Pract 1997;11(1):43-46.

Qudah FJ, M Brannon. Population-based care management. Qual Manage Health Care 1996;5:29-41.

Robinson JR, TK Young, LL Roos, DE Gelskey. Estimating the burden of disease: comparing administrative data and self-reports. Med Care 1997;35(9):932-947.

Roblin DW, PI Juhn, B Preston, R della Penna, S Feitelberg, A Khoury, J Scott. Care coordination for working age adults with multiple chronic medical conditions: targeting patients and estimating potential impact. Unpublished manuscript, 1997.

Shen J, A Iversen. "PACE: A capitated model towards long-term care." Henry Ford Hosp Med J 1992;40:41-44.

Warren BH, T Puls, P Fogelstrom-DeZeeuw. Cost-effectiveness of care management: experiences of a university managed health care organization. Am J Med Qual 1996;11:173-178.

Watt S, J Roberts, G Browne, A Gafni. Age, adjustment, and costs: a study of chronic illnesses. Soc Sci Med 1997;44(10):1483-1490.

Weiner JP, B Starfield, D Steinwachs, L Mumford. Development and application of a population oriented measure of ambulatory care case-mix. Med Care 1991;29:452-472.

Table 1. Descriptive Statistics

A. Expense Distributions	Actual	Predicted (GRAM)
Mean	$1,566	$1,560
Standard Deviation	$5,265	$2,237
Percent with Expense > 0	85.8%	100.0%
Maximum	$277,015	$64,347
Minimum	$0	$20
Skewness	15.21	7.85
Kurtosis	412.73	181.68
Percentiles		
0.99	$19,750	$10,078
0.95	$6,253	$5,592
0.90	$3,268	$3,637
0.75	$1,151	$1,771

B. Demographics	Proportion
Female	0.531
Medicaid	0.005
Medicare Aged	0.103
Medicare Disabled	0.003
Institutionalized	0.004
Spouse in subscriber unit	0.636
Mean Age (also Median)	35.0
Age Range	0 – 103
Interquartile Range	17 – 49

Table 2. Proportion of Actual High-Cost Dollars Correctly Predicted

Percentile Thresholds		GRAM (%)		Prior-Exp. (%)		Logistic (%)	
Actual	**Pred.**	**$ Ratio***	**Sens.****	**$ Ratio**	**Sens.**	**$ Ratio**	**Sens.**
25	**25**	71.2	57.7	67.8	58.1	71.0	58.4
	10	48.0	30.6	41.3	26.8	47.7	31.2
	5	32.9	17.2	27.4	14.4	32.0	17.7
10	**25**	76.8	70.5	71.4	66.4	76.7	70.3
	10	55.2	45.8	47.1	39.2	55.5	45.9
	5	39.3	29.6	32.5	23.4	39.1	30.4
5	**25**	80.3	77.3	74.0	70.6	80.7	77.0
	10	60.0	54.1	51.0	45.4	60.7	54.6
	5	43.7	36.7	36.9	30.6	44.1	37.8

*$ Ratio = total actual expense associated with correct high-cost predictions divided by total actual expense associated with all high-cost cases

**Sensitivity = total number of correctly predicted high-cost cases divided by total number of high-cost cases in the sample

Table 3. Bootstrap Validation Statistics (Percentages) for the Global Risk-Assessment Model

Sensitivity	Predicted		
Actual	**5%**	**10%**	**25%**
5%	36.7 (35.5, 37.7)*	54.1 (52.9, 55.1)	77.3 (76.1, 77.9)
10%	29.6 (28.8, 29.9)	45.8 (45.0, 46.3)	70.5 (69.6, 71.0)
25%	17.2 (16.9, 17.2)	30.6 (30.1, 30.8)	57.7 (57.0, 58.0)

Specificity	Predicted		
Actual	**5%**	**10%**	**25%**
5%	96.7 (96.6, 96.7)*	92.3 (92.3, 92.4)	77.7 (77.7, 77.9)
10%	97.7 (97.6, 97.8)	94.0 (93.9, 94.0)	80.0 (79.9, 80.2)
25%	99.1 (99.0, 99.1)	96.9 (96.7, 96.9)	85.9 (85.7, 86.1)

*Point Estimate (CI 5%, 95%)

Table 4. 15 Most "Sensitive" Disease Groups – GRAM (4,950 (5%) actual high-cost cases / 4,950 (5%) high-cost predictions)

C-B/CRI Variable (Expected CRI are most resource intensive (1) except where shown)	**Coeff. ($)**	**Persons**	**Sens.**	**Spec.**	**PPV**	**NPV**	**Total Acc.**
Chronic, deterioration expected, internal	20,065	110	1.00	0.00	0.68	N/A	0.68
Other serious illness, hospitalization less common	16,456	92	1.00	0.00	0.60	N/A	0.60
Long-term total/supervisory care probable	4,354	67	1.00	0.16	0.40	1.00	0.46
Serious congenital and neonatal disease	7,789	57	1.00	0.00	0.46	N/A	0.46
Complications of medical device, transplant, etc.	11,409	47	1.00	0.00	0.57	N/A	0.57
Serious malignancy	10,701	155	0.98	0.06	0.56	0.67	0.56
Serious microorganism-related, hosp. Common:2	3,231	106	0.97	0.61	0.52	0.98	0.72
Carcinoma in situ, e.g., skin cancer (except serious)	2,846	61	0.94	0.54	0.46	0.96	0.66
Serious microorganism-related, hosp. Common	5,895	168	0.94	0.06	0.41	0.60	0.42
Chronic, continued care required, internal	4,132	465	0.93	0.15	0.41	0.79	0.45
Pregnancy (continuing)	6,081	812	0.92	0.17	0.46	0.74	0.50
Chronic, deterioration expected, systemic	2,797	167	0.89	0.50	0.41	0.92	0.61
Serious non-microorganism-related, hosp. Common	2,221	963	0.86	0.34	0.40	0.83	0.52
Other chronic, systemic and internal	2,686	83	0.86	0.59	0.43	0.92	0.66
Serious malignancy:2	5,998	118	0.85	0.04	0.38	0.30	0.37

Table 5. 15 Most "Sensitive" Disease Groups – Prior-Expense (4,950 (5%) high-cost cases / 4,950 (5%) high-cost predictions)

C-B/CRI Variable (Expected CRI are most resource intensive (1) except where shown)	**Coeff. ($)**	**Persons**	**Sens.**	**Spec.**	**PPV**	**NPV**	**Total Acc.**
Chronic, deteriorating, internal	20,065	110	0.82	0.49	0.78	0.57	0.72
Other serious, hospitalization less common	16,456	92	0.82	0.51	0.71	0.66	0.70
Long-term total/supervisory care probable	4,354	67	0.79	0.65	0.56	0.85	0.70
Serious congenital and neonatal	7,789	57	0.89	0.61	0.71	0.86	0.74
Complications of medical device, transplant, etc.	11,409	47	0.93	0.35	0.66	0.78	0.68
Serious malignancy	10,701	155	0.92	0.43	0.66	0.81	0.70
Serious microorganism-related, hosp. Common:2	3,231	106	0.84	0.58	0.47	0.90	0.66
Carcinoma in situ, e.g., skin cancer (except serious)	2,846	61	0.89	0.56	0.46	0.92	0.66
Serious microorganism-related, hosp. Common	5,895	168	0.81	0.51	0.53	0.79	0.63
Chronic, continued care required, internal	4,132	465	0.66	0.66	0.55	0.76	0.66
Chronic, deteriorating, systemic	2,797	167	0.66	0.77	0.53	0.85	0.74
Serious non-microorganism-related, hosp. Common	2,221	963	0.87	0.42	0.43	0.86	0.57
Other chronic, systemic and internal	2,686	83	0.77	0.72	0.50	0.72	0.74
Serious malignancy:2	5,998	118	0.73	0.51	0.51	0.74	0.60
Other congenital	1,798	169	0.73	0.66	0.44	0.66	0.68

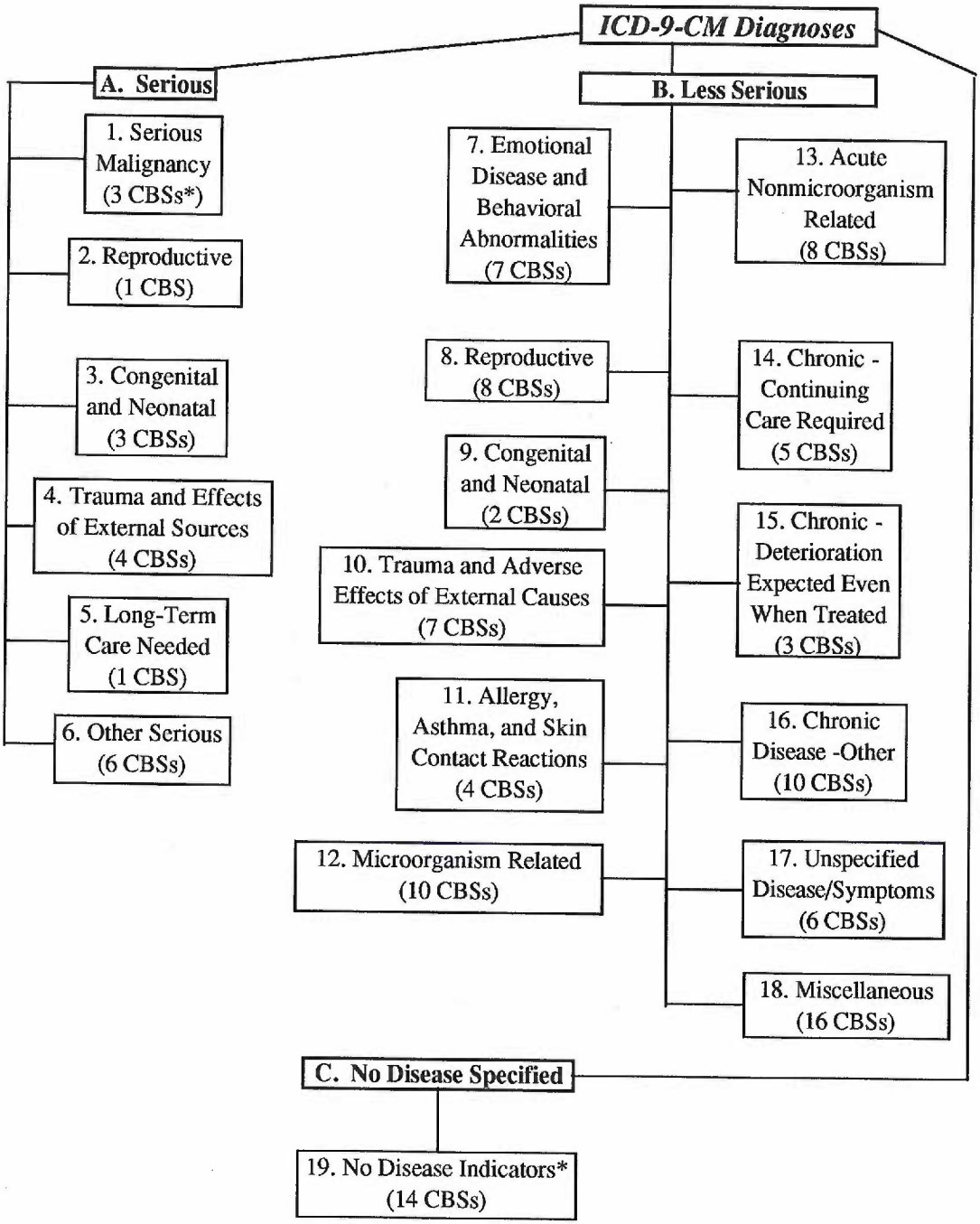
Table 6. 15 Most "Sensitive" Disease Groups – Logistic (4,950 (5%) actual high-cost cases / 4,950 (5%) high-cost predictions)

C-B/CRI Variable (Expected CRI are most resource intensive (1) except where shown)	**GRAM Coeff.**	**($) Persons**	**Sens.**	**Spec.**	**PPV**	**NPV**	**Total Acc.**
Chronic, deteriorating, internal	20,065	110	0.99	0.11	0.71	0.80	0.71
Other serious, hospitalization less common	16,456	92	0.86	0.27	0.64	0.56	0.62
Long-term total/supervisory care probable	4,354	67	0.92	0.40	0.46	0.89	0.58
Serious congenital and neonatal	7,789	57	0.77	0.19	0.44	0.50	0.46
Complications of medical device, transplant, etc.	11,409	47	0.96	0.55	0.74	0.92	0.79
Serious malignancy	10,701	155	0.94	0.13	0.57	0.64	0.57
Serious microorganism-related, hosp. Common:2	3,231	106	0.84	0.70	0.55	0.91	0.75
Carcinoma in situ, e.g., skin cancer (except serious)	2,846	61	0.89	0.58	0.47	0.93	0.67
Serious microorganism-related, hosp. Common	5,895	168	0.93	0.47	0.55	0.90	0.66
Chronic, continued care required, internal	4,132	465	0.87	0.30	0.44	0.79	0.52
Pregnancy (continuing)	6,081	812	0.98	0.02	0.44	0.53	0.43
Serious non-microorganism-related, hosp. Common	2,221	963	0.83	0.49	0.45	0.85	0.61
Chronic, deteriorating, non-internal	1,950	307	0.76	0.57	0.44	0.84	0.63
Serious malignancy:2	5,998	118	0.85	0.21	0.43	0.66	0.47
Chronic, deteriorating, internal:2	2,395	1,847	0.74	0.54	0.39	0.84	0.60

Table 7. 15 Most "Sensitive" C-B/CRI Disease Groups in GRAM and Representative Diseases

C-B/CRI Variable (Expected CRI are most resource-intensive (= 1) except where shown)	**Representative Diseases**
Chronic, deteriorating, internal	Parkinson's disease, angina pectoris
Other serious, hospital not required	Tuberculosis, chronic persistent hepatitis
Long-term care probable	Alzheimer's disease, amyotrophic sclerosis
Complications of medical devices or transplants	Pacemaker malfunction
Serious congenital and neonatal	Cystic fibrosis, low birthweight
Serious malignancies (CRI - 1)	Pancreatic cancer, liver cancer
Serious microorg.; hospital common (CRI - 2)	Botulism, bacterial meningitis
Carcinoma in situ	Non-serious basal/squamous cell cancer
Serious microorg.; hospital common (CRI - 1)	Botulism, bacterial meningitis
Chronic, continuing care required	Epilepsy, mitral stenosis
Pregnancy (continuing into expense year)	N/A
Chronic, deteriorating, systemic	Diabetes mellitus
Other chronic, systemic and internal	Blood, bone marrow, and endocrine diseases
Serious non-microorg.-related, hosp common	Diabetes/ketoacidosis, AMI, stroke
Serious malignancies (CRI - 2)	Malignancy of uterus, lip

Figure 1. Kaiser Permanente Clinical-Behavioral Disease Classification System



*CBS - C-B diagnostic subclass

Figure 2. Receiver Operating Characteristic Curves - 5% High-Cost Prevalence

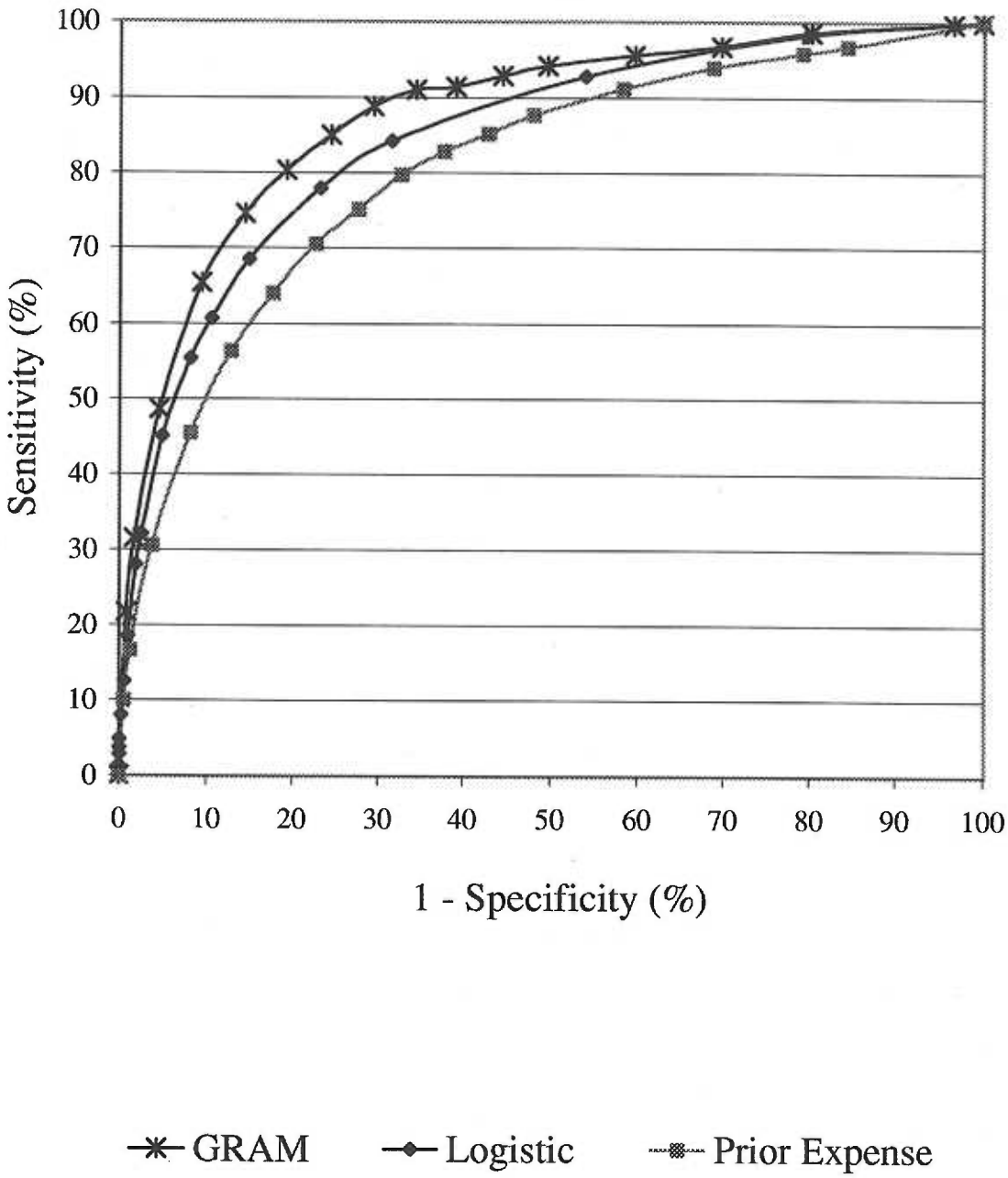


Figure 3. Receiver Operating Characteristic Curves - 10% High-Cost Prevalence

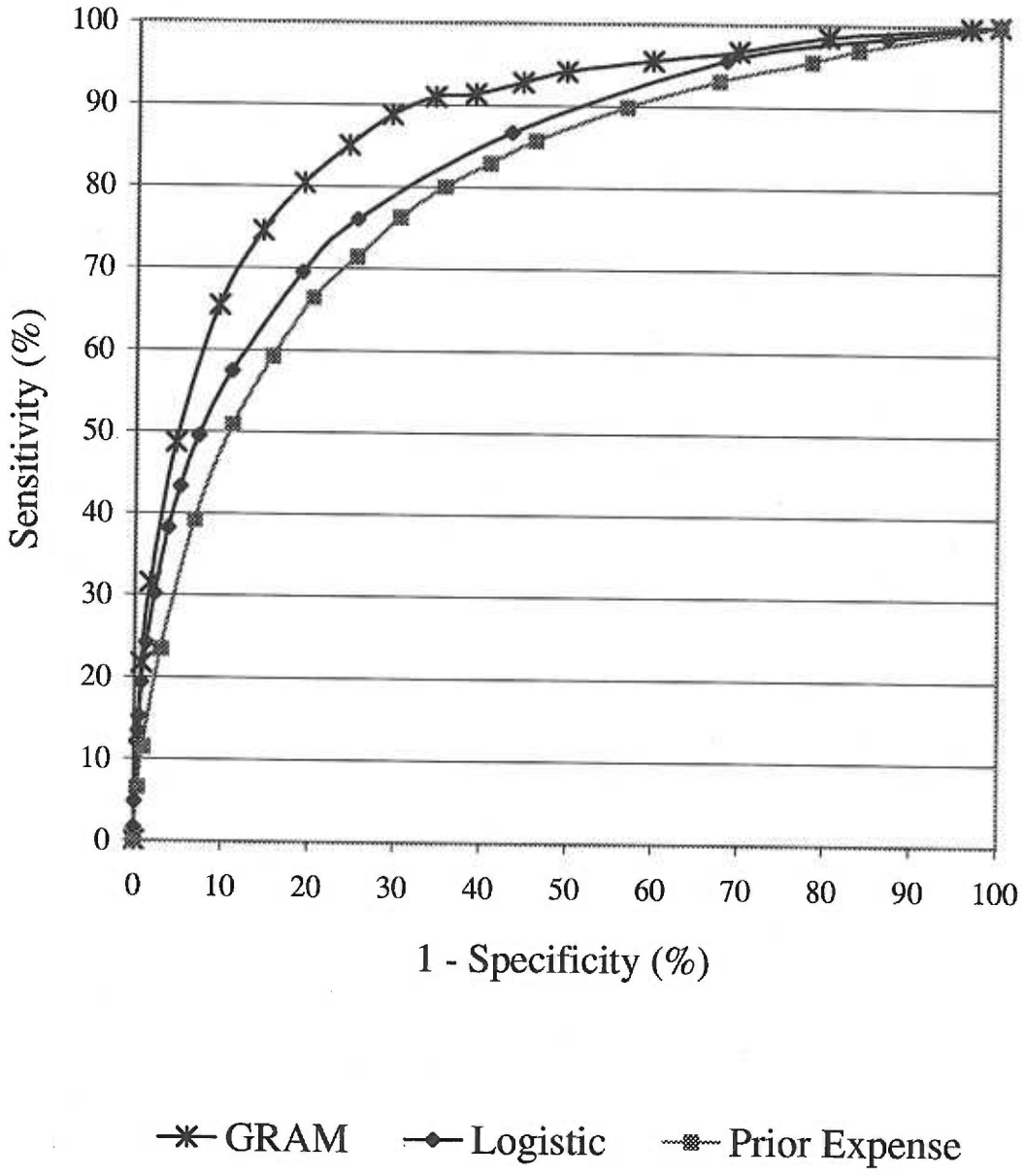


Figure 4. Receiver Operating Characteristic Curves - 25% High-Cost Prevalence