Discussion of "Modeling Mortality Rates for Elderly Heart Attack Patients: Profiling Hospitals in the Cooperative Cardiovascular Project," by Normand, Glickman and Ryan

Stephen E. Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh PA 15213

Martin S. Gaynor
H. John Heinz III School
of Public Policy and
Management
Carnegie Mellon University
Pittsburgh PA 15213

Brian W. Junker*
Department of Statistics
Carnegie Mellon University
Pittsburgh PA 15213

1 Introduction

Normand, Glickman and Ryan (1995, this workshop; hereafter Normand et al.) consider 30-day mortality rates for a retrospectively-studied cohort of 14,581 Medicare patients discharged with acute myocardial infarction (AMI) in 1993 from hospitals located in Alabama, Connecticut, Iowa, and Wisconsin, followed as part of the Cooperative Cardiovascular Project (CCP). A large number of clinical and sociodemographic variables were recorded from individual medical charts and administrative files. Normand et al. produce fitted rates of mortality that are adjusted for patient mix, geography and technological/structural features of the hospitals in each State. The goal is to provide hospital-level profiles of rates of mortality, which might be used as part of an ongoing effort to improve quality of care in the nation's hospitals.

Their analysis can be seen as a determined attempt to improve upon the regression z-scores which had been used to profile hospitals by the Health Care Finance Administration

The order of authorship is alphabetical. We would like to thank Sharon-Lise Normand for conversations that were useful in preparing this discussion.

*Correspondence and preprints.

(HCFA) until 1994, when use of mortality as the single criterion of quality of care was discontinued, and multidimensional process-based measures began to be considered. We see the Normand et al. analysis as a useful first step in replacing inadequately adjusted single-score hospital profiles with more appropriate measures of health care quality, but it is only a first step. (For a non-Bayesian approach to this problem, see Silber et al., 1995).

In our discussion we begin with a brief review of the social and policy backdrop against which the Normand et al. study takes place, which includes increasing demands for accountablility in the health care industry and the evaluation of health care quality by profiling outcomes, including those considered by Normand et al. A proper discussion of the evaluation of outcomes however, requires that we consider carefully what we are trying to measure and why. An appropriate lens through which to view the relationship between the HCFA and the hospitals is that of Agency Theory, and we digress to give a sketch of this economic theory and how it structures the problem at hand. This discussion leads naturally to a consideration of (formal or informal) causal models underlying the health care outcomes of interest. Finally we review the statistical methodology of Normand et al. and suggest some alternatives to the specific goals of their study. In particular, we suggest that with more careful attention to the imputation of missing data, a complete data set—or, in the spirit of multiple imputation, five of them!—might be created with which principals (i.e., HCFA and other insurers and insurance administrators) and agents (hospitals, physicians and other health care providers) in the agency relationship can conduct their own analyses, tailored to the contracts they enter into and the utilities they carry with them.

2 What do we want to measure? Why?

2.1 Background

The past 15 years have seen an increasing demand for accountability in the health care industry. Medicare's Prospective Payment System (PPS hereafter), adopted in 1983, put serious financial pressure on hospitals to take measures to control the cost of care. Hospitals responded strongly by reducing inpatient length of stay and shifting patients (where profitable and clinically feasible) to outpatient settings (not subject to prospective payment) (Cutler, 1995; Guterman and Dobson, 1986; Prospective Payment Assessment Commission, 1990). Other insurers, including private plans and other public programs (Medicaid, State programs, the Veterans' Administration) have followed suit with a variety of payment mechanisms and administrative controls designed to control utilization and costs. The success of PPS in reducing inpatient lengths of stay and controlling costs raised concerns about whether hospitals were achieving those goals at the expense of reduced quality of care (Kahn, 1990; Rogers et al., 1990). Some evidence indicated that patients were being discharged "quicker and sicker," and that adverse outcomes could be a result of these responses to incentives (Kahn, *op cit.*, Rogers *op cit.*, Staiger and Gaumer, 1990). These concerns have been echoed with respect to other insurers' cost containment efforts. Cost containment has been, and continues to be, the major issue confronting health care. However, insurers have begun to demand more comprehensive accountability from providers. In particular, insurers have shown increasing interest in generating measures of provider-specific quality of care.

Medicare again led early efforts in this area. Hospital mortality rates were published until recently, when they were discontinued because of protests over lack of adequate adjustment for patient severity. Nonetheless, concerns over adverse incentives contained in cost control efforts have led public and private insurers to continue efforts to monitor quality. The Cooperative Cardiovascular Project (CCP) discussed in Normand et al. is an example of one effort of this sort. Extensive efforts are being devoted to developing direct measures of inpatient quality of care, and to developing "profiles" of providers' (hospitals and physicians) utilization, outcome, and cost patterns.

2.2 Social Goals

What is lacking from most of these quality-of-care and profiling efforts is explicit discussion of the purpose for developing and collecting these measures, and a design for how these measures will be implemented to achieve these goals. It may seem obvious that collecting data on measures of quality is important. After all, isn't more information a good thing? The answer is, "not necessarily." *Complete* information is always a good thing, but complete information is more of a theoretical benchmark than a feature of the real world. When information is incomplete, more information which is only partial could well be misleading (dranove and Stterthwaite, 1992). Quality in health care is multidimensional and very complex; 30-day mortality rates for AMI (acute myocardial infarction) are unlikely to be a sufficient statistic for hospital quality, even in the limited context of AMI care, and if we act as if they are we might get into trouble.

This is a nontrivial issue, because health care providers respond strongly to incentives and if we blindly penalize hospitals for AMI mortality that exceed some (conditional) expected value, then hospitals have an incentive to do what they can to avoid patients who are more severely ill, a practice sometimes called "patient dumping*". This is probably not what we want.

Let us first suggest at least a general framework for thinking about social goals in health care and then discuss design issues for achieving those goals and how measures of clinical outcomes fit in. One way to state the societal agenda for health care is that its goal is achieving "reasonable" quality at a "reasonable" cost, with some kind of "reasonable" distribution of quantity and quality across the population. A more useful way to think about the organization of health care activities is not as a market for health care *per se*, but as a market for contingent claims. That is, as consumers we each want not health care itself, but the right to purchase the quantity and quality of health care that will maximize our individual utilities should we become ill. This is the insurance contract. Traditionally these contracts have been very incomplete, relying to varying degrees on professionalism, the physician-patient relationship, hospitals' not-for-profit status, and the market, to achieve the desired quantity and quality of care *ex post*. The efforts in recent years to control costs, utilization, and quality are attempts to write more complete contracts *ex ante*.

Thinking about health care as a market for contingent claims emphasizes two points. First, efforts to control costs or quality are part of the insurance contract. Second, *ex ante* contracts will be by necessity incomplete, and optimal contracts should not be "too com-

*Patient dumping is unlikely to be a serious issue in the case of AMI, since hospitals essentially cannot turn away emergency cases; however patient dumping must be considered in the broader health care policy discussion, for it may occur with other less acute conditions.

plete" (see Williamson, 1975, more generally on incomplete contracts). Contracts can not be complete both because of the myriad of possible illnesses and severities and because of the complex and idiosyncratic nature of health care as a product. Health care is in essence a "craft good." That is, the product is customized to the buyer. Indeed, the term "assembly line medicine" carries with it a connotation of poor quality. Further, since there are so many eventualities that we cannot foresee, it is unlikely that a consumer would find it optimal to restrict choices too much *ex ante*. On the other hand, we know that consumers have a difficult time ascertaining product quality, even *ex post*. The trend toward writing more complete contracts specifying price, quantity, and quality *ex ante* represents insurers acting as agents for consumers to specify and monitor the quality of health care that consumers will demand if they become ill. This is not to say that insurers act as perfect agents for consumers. There are agency relations between consumer and insurer, insurer and provider, and provider and consumer.

2.3 Optimal Incentives/Design Issues

The economic theory that informs the analysis of these issues is agency theory (see, e.g., Arrow, 1985; Hart and Holmström, 1987). Agency relationships are defined by three necessary conditions. First, there is comparative advantage between the two parties for the performance of some task. That means that one party is comparatively better at performing the task than the other. Take the case of physicians and insurers. There are two tasks to performed: practicing medicine and underwriting and administering insurance. Trained physicians could potentially both practice medicine and underwrite and administer insurance, but they are comparatively better at the practice of medicine than is the insurer, thus they concentrate on that activity. The argument is similar for the insurer. The implication of the comparative advantage is specialization. Further, if the insurer needs medical treatments to be available to purchasers of an insurance plan, the insurer will contract with the physician to provide these services. The individual performing the task for another party is called the *agent*. The individual hiring the agent is called the *principal*.

The second necessary condition for an agency relationship to exist is divergent objectives.

This means that the objectives of the insurer and the health care provider differ. For example, the insurer may care about quality of care and cost. Health care providers may care about quality of care as well, but they also care about their net income or profits. Since the two parties care about different things, it will not generally be true that the same actions will maximize the objectives of both the principal and the agent. Thus the implication of divergent objectives is that there is a conflict between the two parties.

The third necessary condition is asymmetric information. The fact that there are comparative advantage and divergent objectives don't in and of themselves create a problem. If there is perfect information then the principal knows everything (or alternatively, all the relevant information), including exactly what the agent does. In this case, the principal can simply write a forcing contract for the agent, i.e, one that pays the agent a competitive wage if the desired action is taken, and nothing if it is not. There is unlikely to be perfect information, however. What is more likely is that the agent is better informed about matters relevant to task performance than is the principal. This is called asymmetric information. The implication of asymmetric information is that the principal doesn't know exactly what the agent is doing, nor all the other factors determining performance, so it is not possible to directly observe whether the agent has taken the desired action, nor is it possible to infer it with certainty. Thus, the principal can only infer the agent's action in probability.

Taken together, these three conditions create a relationship between principal and agent where the agent has incentives to shirk at any task and the principal cannot precisely detect that shirking. Further, if the agent is risk averse, it will not be optimal for the principal to fully penalize him for shortfalls in observable outcomes, since they may not be the result of the agent shirking, but rather due random shocks. This means that a contract which is optimal given the information structure will not be fully based on observable outcomes (Ross, 1973).

For example, suppose that a hospital admits patients and produces health outcomes and costs with its use of resources, many of which the insurer cannot observe; however, the actual health outcomes and costs are also the product of patient severity of illness, which is randomly distributed in the population and not observable to the insurer. Thus a poor outcome or high cost might be the result of low effort by the physician or of the patient being severely ill. The insurer is unable to discern which is the source of the poor outcome.

What does the optimal contract look like in this situation? That first depends on whether the relationship between the physician and insurer is temporary or ongoing. First let us consider a temporary one-shot relationship. Agency theory (Holmström, 1979) tells us that the optimal contract: 1) should be increasing in the observable outcome if the probability that the hospital used the optimal amount of resources is increasing in the observed outcome (health or cost); but 2) should not be based purely on the observable outcome if the agent is risk averse. This implies that fixed payments like DRG-based payments (fixed payments for a group of "like" diagnoses called a Diagnosis Related Group), salaries, or capitation are not optimal payment contracts, holding information constant. It also implies that holding providers entirely individually responsible for adverse events is not optimal.

An ongoing, or repeated, relationship alters the information structure and therefore alters the optimal contract between provider and insurer. An ongoing relationship offers repeated observations on the outcome, as opposed to only one observation in a temporary relationship. This offers more precise (statistical) information about the agent's actions than in the temporary relationship. Indeed, if the relationship lasts forever (is repeated an infinite number of times), the perfect information contract becomes feasible (Rubenstein and Yaari, 1983; Radner, 1985; Fudenberg et al, forthcoming). This will be optimal in this situation if agents do not discount the future too much. This result does not necessarily hold if the relationship is repeated, but limited. Thus, ongoing, long-term relationships are beneficial. Their byproduct is the production of more information, which allows the writing of contracts with better incentives.

In the framework above, information acquisition is passive. Insurers acquire information about physicians by observing outcomes, and the more chances they have to observe physicians' outcomes the more information the insurers have about their actions. Information acquisition can also be active. Insurers engage in a number of activities to actively gather information about providers: e.g., physician profiling and utilization review. We refer to any active information gathering as *monitoring*. As indicated above, additional information can improve contracting, which will improve performance. Thus, monitoring can clearly be valuable; however, it is also costly. Consequently in an expanded agency theory model with monitoring, there will be an optimal level of monitoring of the agent's activities. That optimal level will be where the gains in performance due to the additional information acquired through additional monitoring are just balanced by the additional costs.

The need for strong incentives diminishes as more information is acquired from monitoring. As mentioned initially, if information (monitoring) is perfect, there is no need for incentives. Insurers offer take-it-or-leave-it contracts: they simply tell physicians what to do, and if they don't do it they receive no payment (Gal-Or, 1995). This is clearly not a realistic case, but it illustrates the point that there is a tradeoff between monitoring and incentives. The more monitoring is chosen the less reliance there will be on incentives. This indeed, does seem to be the strategy being followed by many managed care plans, which use a combination of financial incentives and administrative controls in their contracts with physicians (Hillman et al, 1989; Hillman, 1991).

A number of statistical and policy issues emerge from this economic discussion. They are as follows:

- Any measures of mortality for a single condition are unlikely to be sufficient statistics for overall provider quality.
- Knowing that a provider had "bad" outcomes without knowing why implies that outcome measures should not be weighted too heavily in policies dealing with providers. Thus analyses must ultimately focus on what caused the bad outcomes and whether hospitals can avoid them.
- Since any measure can only incompletely adjust for "justifiable" variation in mortality, heavy emphasis on such a measure by insurers could cause providers to avoid more severe patients.
- Undue emphasis on such a measure can have unintended adverse consequences, such as causing providers to neglect other dimensions of quality.
- The development and use of such measures has a cost, and optimal design can not be considered independent of these costs.

- Since information is costly, it will not be optimal to obtain complete information. The consequence is that the optimal quality, accounting for information costs, will be less than the quality that is actually clinically feasible.

3 The role of a causal model

Normand et al. begin their paper with a careful explanation of the orientation of the statistical profiling enterprise and how it fits into the broader Cooperative Cardiovascular Project and into public policy discussion of how the HCFA should monitor the appropriateness of processes of care. In light of our sketch of agency theory in Section 2, it is clear that a careful discussion of these issues requires that we consider all variables affecting the processes of care, from the underlying science of AMI, to patient demographics, reimbursement practices, broader social policy variables, etc. By processes of care we mean all the behaviors, individual and collective, of the participants in the health care sector. This includes patient choice of providers, provider acceptance or refusal of patients, and determination of treatment. Taken together, the variables underlying the science, and the processes of care, determine health care outcomes, be they the specific mortality measure that is the focus of the present analyses or other more complex ones.

It is useful to think about these issues by trying to construct a causal model for health carre outcomes so that we can consider causal connections among variables (e.g., see Spirtes, Glymour, and Scheines, 1993; and Pearl, 1996) for illustrations and discussion), as represented in directed acyclic graphs (DAG). In a DAG, nodes (circles) correspond to variables or groups of variables, and directed edges (arrows) correspond to known or hypothesized causal links (association with a causal explanation); the absence of an arrow between a pair of nodes means that the variables are conditionally independent given the parent nodes in the graph, a basic condition for lack of a causal relationship. The risk variables used by Normand et al. fit right into this perspective. After all, why adjust for risk if we don't think that the factors somehow are capturing the processes of who ends up at what hospital for treatment and in what condition?

A starting point for discussing such a model is the schematic DAG given in Figure 1, which has seven groups of variables:

(a) socioeconomic and demographic information on patients, including factors such as age, income and insurance coverage.

(b) medical information on patients, including previous medical and treatment history.

(c) treatment information.

(d) information on reimbursement practises for various treatments, including prices faced by patients and by providers.

(e) hospital and doctor specific factors.

(f) outcome variables.

(g) policy variables, both those under the control of the insurers as well as public policies which shape the environment and insurer administrative practices.

The arrows in the diagram going from one group of variables to the others represent possible causal connections; where we have drawn arrows going back and forth between two groups of variables we imagine that there are one-way causal links in both directions between individual variables in the two groups. We are not experts on the medical science and processes of care issues surrounding AMI's, and thus we have chosen not to be specific about exactly which variables should be listed in each group, but we do note that Normand and her colleagues have given us a good start. The details do matter, however.

Having spelled out the possible causal structures in this way, we can now ask how the models used to develop profiles and risk-adjustments fit with the overall causal structure. For example, do the marginal calculations in the Normand et al. paper, which in effect average over all other variables and factors, adequately represent what we need to know about the actual processes and their implications? There are two parts to this question: one is theoretical and the other empirical. The theoretical question is whether their final logistic regression model is equivalent to some (marginal) subgraph of Figure 1. The second question is whether the model provides a reasonable fit to the data and whether, even in the face of

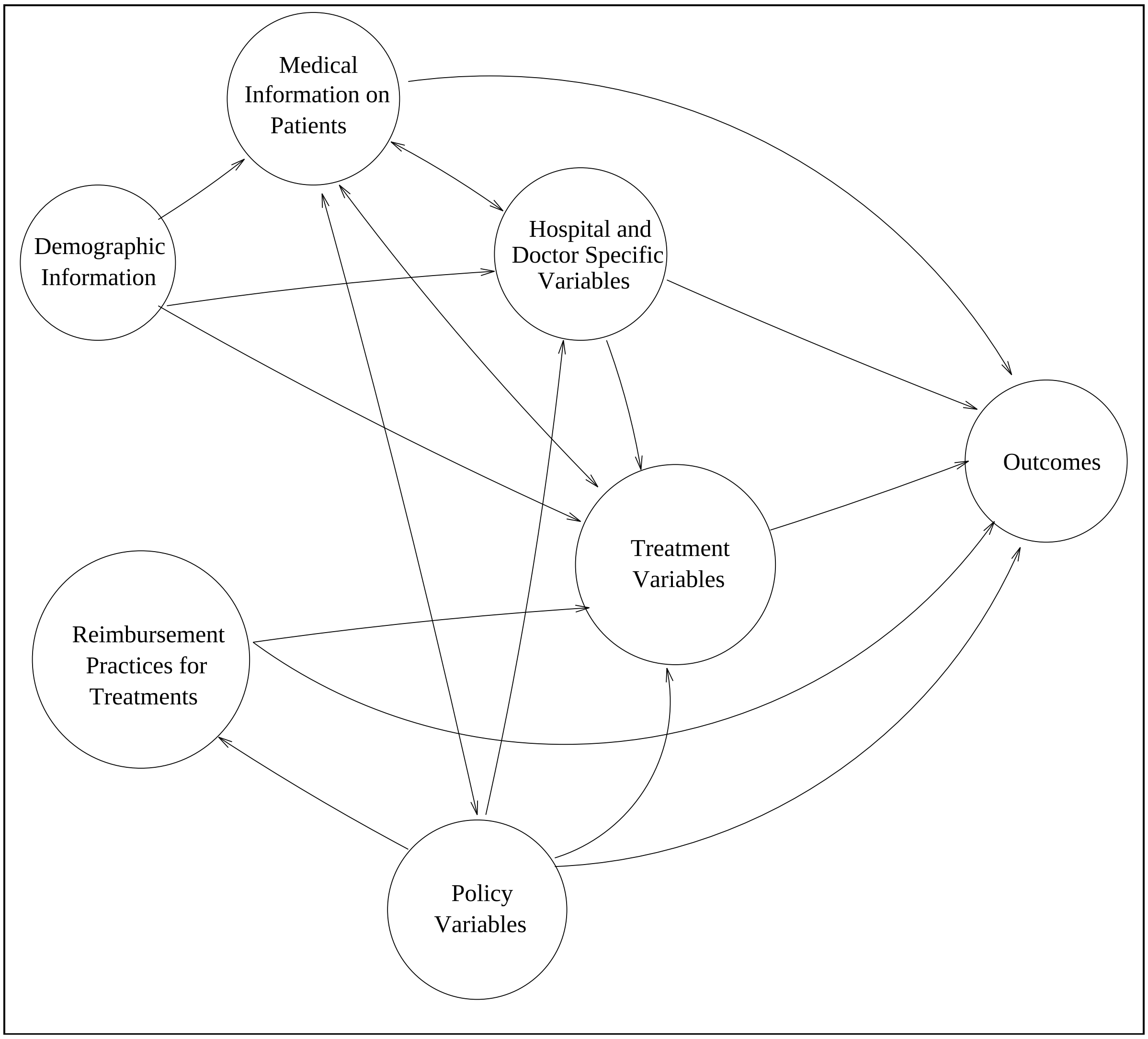

Figure 1: A plausible causal model for the discussion of quality of care.

an answer of "no" to the theoretical question, the empirical logistic regression results are robust to misspecification. We return to these themes in Section 4.

Which bring us to the final point of this causal diagram: What are the implications of various policy responses to the analyses? Such policy variables may actually lead to changes in treatment and other variables and even to a different set of causal relationships.

4 Analyzing the data

4.1 Brief Review of the Normand et al. Analysis

The data considered by Normand et al. from the CCP lend themselves naturally to analysis with a hierarchical Bayes model that nests patients within hospitals within hospital cohorts (teaching vs. nonteaching, Wisconsin vs. Alabama, etc.). Normand et al. seek to predict fitted rates of mortality from predictors in such a model. The patient-level predictors can be divided into two groups, admission predictors (comorbidities, demographic, and other data collected or believed to be true at the time the patient was admitted) and non-admission predictors (collected or imposed as the result of post-admission health care, such as adminstration of drugs, revascularization procedures, etc.). The hospital predictors included geographical location, relevant specialized services, size, and the existence of teaching and/or residency programs. Normand et al. model individual hospital effects on AMI mortality as random effects within cohorts determined by the hospital level predictors. Aside from the sheer number of observations (14,581 Medicare patients), the principal challenges faced by Normand et al. are imputation of missing data and variable selection from the approximately 49 predictors they consider.

4.2 Divide and Conquer

Normand et al. pursue a "divide and conquer" strategy for the analysis, breaking the stastistical analysis problem up into three steps (Steps 2–4 of Normand et al.), and they use different probability models and modes of inference at each stage in the analysis. In Step 2, they impute missing data using the General Location (GL) model (Liu and Rubin, 1995), yielding 5 complete data set draws $(y,x,w)_1$, $(y,x,w)_2$, $(y,x,w)_3$, $(y,x,w)_4$, and $(y,x,w)_5$ from $p(y,x,w|y,x_{obs},w_{obs})$, where y represents the outcome, 30-day mortality, x represents patient level covariates and w represents hospital-level covariates. In Step 3, they perform variable selection and transformation by applying S-PLUS's classification trees (CT) and generalized additive models (GAM) functions to the five imputed data sets. Finally in Step 4, they fit a Hierarchical Bayes (HB) model of the form

$$\begin{aligned} \textbf{Stage I}: &\quad \text{logit}(P[Y_ij = 1|\underline{x}_{ij}, \underline{\beta}_i]) = \underline{x}_{ij}\underline{\beta}_i \\ \textbf{Stage II}: &\quad \underline{\beta}_i|\Gamma, D, \underline{w}_i \sim N(\Gamma\underline{w}_i, D) \\ \textbf{Stage III}: &\quad \Gamma \sim d\Gamma, \quad D \sim dD \end{aligned} \tag{1}$$

where D has block diagonal structure so that the error terms for different hospitals are independent and homoscedastic. In each step only additive models were considered, for computational convenience (for example, interactions would have had to be coded by hand in S-PLUS's CT and GAM routines).

Breaking up the analysis in this way and applying different models at each step immediately inspires concerns: Was the HB model used for inference consistent with the GL model used for imputation? Did the additive multinomial CT model and the additive GAM model provide adequate approximations to the marginal probability distribution $p(y,x,w)$? How much did the additivity assumptions made in each step of the analysis affect the outcomes of subsequent steps—Might different variables have been chosen for the final model, had interactions been coded in the earlier steps? What is the effect, in the two variable selection phases, of ignoring the hierarchical structure of step 4 and and merely pooling all the variables together? What is the effect of omitting from the analysis the specific treatment assigned to the patient, clearly an important predictor of mortality?

Although one might quibble over such concerns, the analysis presented is at least plausible and certainly clear. Indeed such concerns can only be raised because of the clarity with which Normand et al. describe their analysis. Given the size of the problem one is certainly tempted to break it up; and imputation, variable selection and inference steps are natural. Although there now exist various working implementations of formally Bayesian variable selection techniques (Madigan and Raftery, 1994), the size of the variable selection task here suggests that one consider at least some computationally-simpler frequentist methods for initially screening the variables. We can view the Normand et al. analysis as an approximation to a fully Bayesian analysis that treats imputation, variable selection, and inference under the roof of a single probability model for data and parameters. Formal and informal sensitivity analysis absent from the paper would be an effective antidote to concerns over the efficacy of their approximation to the fully Bayesian approach; most of the questions raised in the previous paragraph, as well as the robustness questions raised at the end of Section 3, could have been addressed in this way.

4.3 Variable Selection

The variable selection in step 3 of the Normand et al. analysis leads to a somewhat surprisingly limited set of predictors. The final step 4 model, based on the variable selection, is, in a `GLIM`-like notation,

$$\begin{aligned}\log(mortality) &= 1 + Age + \log(creatitine) \\ &\quad + \log(respiration\ rate) + (mean\ arterial\ pressure)\end{aligned}$$

All of these variables are patient-level variables, and in fact are admission (pre-treatment) variables. It seems remarkable that no non-admission variables (treatment-related) or hospital-level variables make it into the model.

This must have been surprising to Normand et al. also, since despite the fact that no hospital-level predictors were selected by the CT procedure, they still included dummies for the States (Alabama, Connecticut, Iowa, Wisconsin) in the hierarchical model. Indeed, while examining model-fit at the second stage, Normand et al. remark that posterior predictive distribution checks suggest that two other hospital-level predictors, existence of a residency program and absence of open-heart surgery facilities, seem to be hospital-level predictors that were missed by the variable selection.

Clearly, the variable selection in the marginal distribution $p(y, x, w)$ was not very sensitive to the second-stage covariates w. Some crude simulations, not reported here, make clear that this isn't always so: CT variable selection is able to pick up second stage covariates in data generated from HB models similar to that used by Normand et al., if the values of the second-stage covariates are well distributed among observations. On the other hand many of the hospital-level predictors don't show much variability among the hospitals studies in the CCP. For example, Intensive Care Units are absent at only 4.2% of the hospitals in this CCP data set. Thus it is unsurprising that they do not show up as important predictors of mortality, in this study.

The exclusion of some hospital-level variables early in the analysis, viz. treatment, appears to be at least partly a political decision, based on past experience with acrimonious relationships that developed when the HCFA's Peer Review Organizations directly reviewed random samples of individual hospitalizations in profiling hospitals. As outlined in Section 2.3, a crucial component of the principal/agent relationship is monitoring, which is improved as the principal learns more about matters relevant to task performance by the agent. Given the *prima facie* case for treatment variables as influencing outcome, not to mention exonerating hospitals from inappropriate responsibility for poor outcomes with extremely ill patients, it is hard to understand why they were left out of the variable selection and modeling phases of the Normand et al. study. While the goal may be to develop models that do not require them, we think that it is almost certain that we will be able to understand their value more clearly by initially collecting and analyzing detailed data that include treatment choices, etc. In this way we can establish whether treatment variables are in fact important for monitoring, as well as for subsequent modifications of the principal/agent contract.

A somewhat more formal approach to variable selection, related to the process by which Normand et al. added back in second-stage variables after initially screening them out with non-hierarchical CT and GAM models, would be to observe that in the HB model (2), the responses y are independent of the hospital-level predictors w, given the patient-level predictors x and the Stage I regression coefficients β. Thus the form of partial exchangeability determined by x in the marginal distribution of y, averaging over parameters, is the same whether we condition on x or on (x, w). In particular, under some mild assumptions, although the magnitudes of the coordinates of β may change, whether or not a β coordinate is zero should not change, under the model, whether or not we also condition on w. Thus one could do variable selection for the patient-level variables, and then separately consider variable selection for the hospital-level variables, using a method that can asssess whether the partial exchangeability structures induced by including each hospital-level variable makes sense. This also suggests criteria for considering a reduced model that deliberately averages over some covariates, as discussed toward the end of Section 3.

4.4 The Fitted Inferential Model

Normand et al. report posterior means and standard deviations of the second stage parameters from the HB model fits, using Markov Chain Monte Carlo (MCMC) in Table 1.7. They display corresponding maximum likelihood estimates and standard errors for the first stage model fitted separately in each of the four States in Table 1.8. The tables are strikingly similar. This is to be expected, since the Stage III priors were flat and disallowed correlation between hospitals—see equation (2)—so that, in principle, the Bayesian posterior modes should be exactly equal to the ML estimates. In this sense, the ML estimates provide a check on the MCMC estimates; had the two sets of estimates been vastly different this would have suggested either a problem with the algorithms or large amounts of skew in the posterior distribution of the Stage II parameters.

Figures 6, 7, and 8 of Normand et al. summarize the estimated posterior mean rate of mortality for a randomly selected patient from each hospital, adjusted for the four patient-level predictors and the State dummies in the HB model. Figure 6 is especially interesting. For each State we see the standard Bayesian shrinkage plots, with the unadjusted sample mortalities at the top of each subplot connected by a line segment to the corresponding adjusted, posterior mean mortalities at the bottom of each subplot. In all States but Connecticut, most hospitals' mortality rates are shrunk toward State means, with one or two outlier hospitals really "de-shrunk" away from the pack. Connecticut has an outlier of a different sort, a hospital that was way out before shrinkage, and was fully incorporated "into the pack" after shrinkage. If additional hospital-level predictors had been included in the model, we might expect to see these outlying hospitals shrunk towards second-stage means determined by the new predictors.

4.5 The Role of High-Stakes and Hospital-Level Variables

What causes the outliers in Figure 6? What causes there to be heterogeneity in the risk of mortality within and among the four States considered? Surely State is a surrogate for some more informative covariate(s). Normand et al. suggest heterogeneity in data abstraction as a possible source of the problem, and this should certainly be investigated. However, in view of our discussion in Section 3, it seems clear that many variables that are related to outcome have been omitted from the Normand et al. analysis.

Normand et al. remind us that inclusion of treatment information in the model is clearly a sensitive issue, for at least two reasons. First, treatment information carries with it very high stakes for individual doctors, as well as hospitals, who can be held accountable for treatment decisions, and in the past the analysis of individual hospitalizations has caused excessive friction between doctors/hospitals and HCFA Peer Review Organizations. Second, inclusion of many different hospital-level variables (treatment, or at least standard-of-care, being one such) creates an adjustment structure in which there are not merely four "standards", one for each State, but rather many national "standards" depending on which mix of hospital-level variable values each hospital possesses.

Three important points bear emphasis here. First, our discussion of agency theory in Section 2 makes clear that analysis of a narrowly-defined outcome alone is insufficient for establishing adequate contracts between principal (HCFA) and agents (hospitals). Additional information on the processes of care reduces asymmetry of information and implies that less strong incentives are needed in contracts. In particular with the use of more information it may be possible to encourage improvements in health care practices without unintended side-effects such as patient-dumping.

Second, as suggested in Section 4.3, though the aim of the modeling enterprise may be to create models in which these high-stakes variables are not present, it is not safe to proceed by simply ignoring them. Otherwise one is integrating over the (empirical) distribution of a potentially important covariate, that is unobserved by fiat, and inducing dependence not accounted for in the HB model of this paper. If, as suggested in Section 4.3, certain conditional independence relationships hold between the observed and unobserved data, then no harm is done. Otherwise, how this dependence plays out in the analysis is impossible to judge without considering including these sensitive covariates.

Third, the enterprise of producing adjusted estimated rates of mortality is precisely the enterprise of creating different standards for different risk groups. To their credit, Normand et al. have suggested a common measure that allows comparison across the risk groups, by considering measures of outlying risk that are relative to the adjusted groups. For example in their Figure 8, quite a different intervention might be suggested by the information that a particular hospital is an outlier in the distribution of probability of excessive risk among hospitals in the same State, than by the information that overall mortality is higher in Alabama than Connecticut.

4.6 Imputation Revisited

Returning to some of the themes of Section 2, each of the participants in the health care industry—broadly: providers, insurers and consumers—evaluates health care outcomes with a different utility function, that may even suggest that different (sets of) outcome variables are important to different participants. Although identifying hospitals with outlying estimated mortality rates is a first-order description of where further analysis and possible intervention are required (the losers in the adjusted-mortality horserace), it is only a first step in the larger problem of monitoring and improving health care quality.

Given that each actor brings a different utility and focus to the problem, the most useful general contribution a statistician can make to the problem may be to provide completed data sets, by imputation, and perhaps broadly-usable estimates of several important outcome variables (surely mortality rate would be among them). Normand et al. have given us a path to follow in attacking this larger imputation problem. Their HB model is a place to start in building an imputation model that—unlike the General Location model of Liu and Rubin (1995)—unambiguously reflects the nested structure of patients and doctors within hospitals, hospitals within States, etc. A plausible next step would be to expand their HB model, to include more predictors as sketched in Section 3, and use it for both (a) imputation of missing values and (b) estimation of adjusted mortality rate and other outcome variables.

The imputation should involve as many plausible predictors as is computationally feasable, including treatment informtion where available, so that subsequent analyses based on the imputed data can condition on any plausible predictor of mortality or other outcomes. Johnson, Mislevy and Thomas (1994) outline the pitfalls of omitting covariates in an imputation model whose output will be open to general use, and they show that very many covariates indeed—about 150—can be practically incorporated into the imputation model. Multiple copies of the data set (multiple imputations for the missing data with subsequent estimates of the collection of outcome variables) might be made available to anyone with a legitimate interest in health care quality monitoring, together with instructions for producing approximately-posterior analyses of the outcome variable(s), using standard statistical packages.

The problem that lead to the work of Johnson, Mislevy and Thomas (1994) is a general-purpose imputation project in a different substantive context, the National Assessment of Educational Progress (NAEP). A general description of the NAEP statistical analyses is given by Special Issues of the *Journal of Educational Statistics* (1992, No. 2, pp. 93–232; see Zwick, 1992, for an introduction) and *Journal of Educational Measurement* (1992, No. 2, pp. 93–194; see Algina, 1992, for an introduction). Although the substantive context is different, the the goal of producing general-purpose data sets completed by imputation for use by a wide range of policy analysts is the same, and NAEP and its contractors and subcontractors have been remarkably successful in achieving this goal. Thus there is hope that this could succeed in health care quality monitoring also.

5 Conclusions

The evaluation of health care outcomes, a consequence of the current interest in quality of care and cost containment, occurs in many different contexts, depending on who is doing the evaluation and why. It is essential that discussion of the outcomes of evaluation be couched in sensible economic and behavioral models of the health care sector.

Profile models, such as those leading to the z-scores formerly used by the HCFA, as well as the vast improvements on this methodology offered by Normand et al. in the present article, are a subset of the larger class of models that relate outcome variables to medical science, specific course of treatment, demographics, and policy variables. The schematic causal model discussed in Section 3 is an attempt to illustrate this larger class of models. In terms of the economic theory of agency discussed in Section 2, adjusted profile models can be thought of as arising out of the defense of agents against careless analysis of a single outcome like 30-day mortality: "My hospital may have a high mortality rate *but...*". We feel that the profiling enterprise should be replaced with evaluation of health care outcomes in the context of the broader models discussed in Section 3, aided by economic theory as discussed in Section 2.

Normand et al. have shown how far the profile methodology can be carried by careful and tireless data analysis, and great expertise in the vast array of modern statistical methodologies available. Their analysis can be viewed as a reasonable first approximation to a fullly-Bayesian analysis in the large and complex CCP data set.

An advantage of a fully-realized, unified Bayesian treatment is that it leads naturally to broader, general-purpose imputation models. Analysis of the processes of care, even for a single illness such as AMI, is far too large a problem for a one-off data analysis. Using the Normand et al. analysis as a starting point, we see an opportunity to generate completed data sets from an imputation model that accounts for the hierarchical structure of the problem, and incorporates as many potentially-relevant variables/covariates as is technically feasable.

An analogous general-purpose imputation strategy, within a Bayesian framework, has been developed for the National Assessment of Educational Progress. The NAEP imputation effort is also aimed at providing data sets completed-by-imputation that can be used by any analyst of education and U.S. educational policy. The experience gathered in the NAEP imputation effort should be combined with Normand et al.'s specific health care expertise to produce compete data sets that would be valuable to a vast array of health care analysts, from providers to insurers to consumers.

A final issue is economic. The issues of outcomes, monitoring, incentives, and design, cannot be considered independently of the costs of implementing the design. Extensive data-gathering efforts like the CCP are expensive and produce a large quantity of unfiltered data. To realize benefits commensurate with the costs of the CCP and similar projects, we must be able to make effective use of the data, and this usually requires that it be organized and filtered appropriately. Normand et al.'s excellent and clear organization of the problem, and their analyses, will surely reduce some of these costs for future users.

References

Algina, J. (1992). Special issue: the National Assessment of Educational Progress (Editor's Note). *Journal of Educational Measurement, 29,* 93–94.

Arrow, Kenneth J. (1985) "The Economics of Agency," in John Pratt and Richard J. Zeckhauser, eds., *Principals and Agents,* Cambridge, MA: Harvard Business School.

Cutler, David M. (1995). "The Incidence of Adverse Medical Outcomes under Prospective Payment," *Econometrica, 63,* 29–50.

Dranove, David and Mark A. Satterthwaite. (1992). Monopolistic competition when price and quality are imperfectly observable. *Rand Journal of Economics, 23,* 518–534.

Fudenberg, Drew, Holmström, Bengt, and Paul Milgrom (1995). "Short-Term and Long-Term Agency Relationships," *Journal of Economic Theory,* forthcoming.

Gal-Or, Esther (1995) "Optimal Reimbursement and Malpractice Reform in Health Care Markets," manuscript, Katz Graduate School of Business, University of Pittsburgh.

Gaynor, Martin and Paul J. Gertler (1995) "Moral Hazard and Risk Spreading in Medical Partnerships," *Rand Journal of Economics, Winter,* forthcoming.

Guterman, Stuart and Allen Dobson (1986) "Impact of Medicare Prospective Payment System for Hospitals," *Health Care Financing Review, 7,* 97–114.

Hart, Oliver and Bengt Holmström (1987) "The Theory of Contracts," in *Advances in Economic Theory, Fifth World Congress,* Truman Bewley, ed., Cambridge, UK: Cambridge University Press.

Hillman A.L., Pauly M.V., Kerstein J.J. (1989) "How do Financial Incentives Affect Physicians' Clinical Decisions and the Financial Performance of Health Maintenance Organizations," *New England Journal of Medicine, 321, 2,* 86–92.

Hillman, Alan L. (1991) "Managing the Physician: Rules Versus Incentives," *Health Affairs, Winter,* 138–146.

Holmström, Bengt (1979) "Moral Hazard and Observability," *The Bell Journal of Economics, 10, 1,* 74–91.

Johnson, E. G., Mislevy, R. J., and Thomas, N. (1994). Theoretical background and philosophy of NAEP scaling procedures. Chapter 8, pp. 133–146 in Johnson, E. G., Mazzeo, J. and Kline, D. L. (1994). *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading.* Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Kahn, Katherine L. et al. (1990) "Comparing Outcomes of Care Before and After Implementation of the DRG-Based Prospective Payment System," *Journal of the American Medical Association, 264,* 1984–1988.

Liu, C. H. and Rubin, D. B. (1995). *M: Data manipulation environment programming language graphics,* Department of Statistics and Department of Health Care Policy, Harvard University and Harvard Medical School. Cambridge, MA.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window. *Journal of the American Statistical Association, 89,* 1535–1546.

Normand, S.-L., Glickman, M. E., and Ryan, T. J. (1995). Modeling mortality rates for elderly heart attack patients: profiling hospitals in the Cooperative Cardiovascular Project. *Invited paper, this workshop.*

Pearl, J. (1996). Causal diagrams for empirical research. *Biometrika, 82,* (in press).

Phelps, Charles E. (1992) *Health Economics,* New York: HarperCollins Publishers.

Prospective Payment Assessment Commission (1991) *Medicare Prospective Payment and the American Health Care System,* Washington, D.C.: Prospective Payment Assessment Commission.

Radner, Roy (1985) "Repeated Principal-Agent Games with Discounting," *Econometrica, 53,* 1173–1198.

Rodgers, William H. et al. (1990) "Quality of Care Before and After Implementation of the DRG- Based Prospective Payment System: A Summary of Effects," *Journal of the American Medical Association, 264,* 1989–1994.

Ross, Steven (1973) "The Economic Theory of Agency: The Principal's Problem," *American Economic Review, 63,* 134–139.

Rubinstein, Ariel and Menachem Yaari (1983) "Repeated Insurance Contracts and Moral Hazard," *Journal of Economic Theory, 30,* 74–97.

Silber, J. H., Rosenbaum, P. R. and Ross, R. N. (1995). Comparing the contributions of groups of predictors: which outcomes vary with hospitals rather than patient characteristics? *Journal of the American Statistical Association, 90,* 7–18.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, prediction, and search.* New York: Springer-Verlag.

Staiger, Douglas and Gary L. Gaumer (1990) "The Impact of Financial Pressure on Quality of Care in Hospitals: Post-Admission Mortality Under Medicare's Prospective Payment System," unpublished manuscript, Stanford University.

Williamson, Oliver E. (1975) *Markets and Hierarchies,* NY: Free Press.

Zwick, R. (1992). Special issue on the National Assessment of Educational Progress. *Journal of Educational Measurement, 17,* 93–94.