



The Unexpected Influence of Physician Attributes on Clinical Decisions: Results of an Experiment

John B. McKinlay; Ting Lin; Karen Freund; Mark Moskowitz

Journal of Health and Social Behavior, Vol. 43, No. 1. (Mar., 2002), pp. 92-106.

Stable URL:

<http://links.jstor.org/sici?sici=0022-1465%28200203%2943%3A1%3C92%3ATUIOPA%3E2.0.CO%3B2-Z>

Journal of Health and Social Behavior is currently published by American Sociological Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/asa.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

The Unexpected Influence Of Physician Attributes On Clinical Decisions: Results Of An Experiment

JOHN B. MCKINLAY
TING LIN

New England Research Institutes, Inc.,

KAREN FREUND
MARK MOSKOWITZ

Boston Medical Center

Journal of Health and Social Behavior 2002, Vol 43 (March): 92–106

This experiment was designed to determine: (1) whether patient attributes (specifically a patient's age, gender, race, and socioeconomic status) independently influence clinical decision-making; and (2) whether physician characteristics alone (such as their gender, age, race, and medical specialty), or in combination with patient attributes, influence medical decision-making. Methods. An experiment was conducted in which 16 ($= 2^4$) videotapes portraying patient-physician encounters for two medical conditions (polymyalgia rheumatica (PMR) and depression) were randomly assigned to physicians for viewing. Each video presented a combination of four patient attributes (65 years or 80 years of age; male or female; black or white; blue or white collar occupation). Steps were taken to enhance external validity. One hundred twenty-eight eligible physicians were sampled from the northeastern United States, with numbers balanced across 16 ($= 2^4$) strata generated from the following characteristics (male or female; < 15 or ≥ 15 years since graduation; black or white; internists or family practitioners). The outcomes studied were: 1) the most likely diagnosis; 2) level of certainty adhering to that diagnosis; and 3) the number of tests that would be ordered. Results. Patient attributes (namely age, race, gender, and socioeconomic status) had no influence on the three outcomes studied (the most likely diagnosis, the level of certainty, and test ordering behavior). This was consistent across the two medical conditions portrayed (PMR and depression). In contrast, characteristics of physicians (namely their medical specialty, race, and age) interactively influenced medical decision-making. Conclusion. Epidemiologically important patient attributes (which Bayesian decision theorists hold should be influential) had no effect on medical decision-making for the two conditions, while clinically extraneous physician characteristics (which should not be influential) had a statistically significant effect. The validity of idealized theoretical approaches to medical decision making and the usefulness of further observational approaches are discussed.

*Acknowledgements: This work was supported by a grant (AG12437) from the National Institute on Aging (National Institutes of Health). The authors thank the field interviewers, the project team, and the video production staff. We also are grateful to Drs. Henry Feldman and Sonja McKinlay (biostatisticians)

and to Lisa Marceau and Agnes Migliaccio for their many contributions. We also acknowledge the helpful comments of an anonymous reviewer and Michael Hughes. (Address correspondence to John McKinlay at New England Research Institutes, 9 Galen St., Watertown, MA; email: johnm@neri.org)

Studies of the doctor-patient relationship were the almost exclusive domain of medical sociologists from around the mid-1960s through the 1980s (Clark, Potter, and McKinlay 1991). Reflecting the perspective of these disciplines, the medical encounter was viewed as a closed micro social system, encompassing the interaction of individuals with competing and sometimes conflicting backgrounds and interests (Henderson 1956; Freidson 1962; Bloom and Summey 1976). Influenced by the dominance of "the" profession at that time (Balint 1957; Freidson 1970a, 1970b), models of the doctor-patient relationship, while recognizing the social control exercised by physicians (Zola 1972; McKinlay 1973; Waitzkin and Stoeckle 1976), retained the ideal of a "value-free" or "non-judgmental" relationship. Parsons (1951) exemplified this view in his idealized depiction of physicians as "effectively neutral," employing "universalistic criteria" in the assessment and management of particular cases. In the "ideal case," individual patient characteristics (like gender, age, socioeconomic status, race, ethnicity, type of health insurance, and physical attractiveness) should not intrude on the relationship.

Societal reaction or labeling theorists eventually challenged this view by highlighting the often subtle differential influence of medically extraneous patient characteristics (like race, socioeconomic status, and physical attractiveness) on the decision making behavior of mandated labelers (professionals) (Bittner 1967; Kitsuse 1968). Much, but not all, of this work viewed "the" doctor as engaged in medical imperialism and as a powerful agent of social control (Zola 1972; McKinlay 1973). While generally employing weak research methods, their findings probably do reflect what actually occurs in most professional encounters (patients or clients with different social attributes often receive differential treatment over and above the influence of presenting symptomatology).

Epidemiological studies over several decades have identified and estimated the contribution of patient attributes (like gender, race/ethnicity, age, and socioeconomic status) to the distribution (prevalence) of a broad-range of disease groups (now referred to as "social patterning"). Such attributes are often considered to be "risk factors." The best U.S. example is the Framingham Heart Study

(Wilson 1994; Kannel 1995), which identified and estimated the contribution of numerous patient attributes to the prevalence of coronary heart disease (for example, it is reportedly more common with increasing age and among males). Medical sociologists studying the doctor-patient relationship overlook these important epidemiologic findings. They are, however, not overlooked by an emerging group of medical decision theorists (mainly economists, biostatisticians, and physician-trained health services researchers) who focus not on "the relationship," but rather only on the decision-making of physicians. According to these decision theorists, patient attributes that are epidemiologically associated with a medical condition should (ideally) influence decision-making with respect to that condition. A good Bayesian decision-maker is a physician who gives appropriate weight to epidemiologically relevant patient attributes (as apriori probabilities) during the process of making a diagnostic and patient management decision (Weinstein 1980; Elstein, Shulman, and Sprafka 1978; Lustad 1968; Schwartz et al. 1973; Pauker 1982). In other words, symptoms suggestive of, say, coronary heart disease should be interpreted differently in an older male patient compared with a younger female patient with exactly the same symptomatology. Decision-making differences due to social attributes should reflect the social patterning of disease. In contrast to the ideal type encounter described earlier by Parsons and others, it is now believed that key patient attributes ought to influence clinical decisions.

Our focus on the contribution of non-medical influences illustrates some differences between two general approaches to medical decision-making. First, prescriptive decision-making studies (employing a somewhat idealized view) focus on how physicians, as rational actors utilizing abstract statistical reasoning, ought to make appropriate medical decisions. These studies tend to be conducted by statisticians and economists, and they appear "*decontextualized*" with a homo medicus view. They generally overlook the effect on decision-making of the types of influences that social scientists have identified as profoundly affecting the doctor-patient encounter (e.g., attributes of patients, characteristics of providers and the organizational contexts in which encounters occur). Second, descriptive decision-making studies focus on how physicians in the context

of everyday practice actually make their decisions. The origins of this work can be traced to earlier social scientists (especially symbolic interactionists), and it gives weight to the influence of the social and psychological characteristics of each of the actors (namely the doctor and the patient) and the social location of the encounter (e.g., a managed care setting) on the decisions that result.

As discussed, much work has focused on the characteristics and behavior of patients with respect to the clinical encounter. There is recent increased interest in the influence of provider attributes and the organizational contexts within which the encounter occurs (e.g., in profit versus not-for profit structures). This shifting emphasis may reflect the managerial ethos now shaping health care research: The health care industry is motivated to constrain provider behavior (increase productivity and reduce worrisome practice variations), and much health services research focuses on improving clinical appropriateness (through practice guidelines) and the diffusion of evidence-based practices. Employing a descriptive approach, we recognize that, like all human actors, physicians bring to the medical encounter motives and often-unrecognized biases that reflect lifelong socialization in the surrounding culture.¹ Formal training in clinical decision making and the imposition of clinical guidelines are unlikely to completely eliminate the everyday influence of long-held social preferences (or prejudices) concerning, for example, age, race, gender, and socioeconomic status. In other words, over and above the social patterning of some illness condition, the variable behavior of physicians may, not surprisingly, also reflect economic and organizational interests and social prejudices, among numerous other influences. A simplified working model depicting some of these influences on medical decision-making is presented in Figure 1 (McKinlay and Marceau 2001).

Disentangling (or unconfounding) the independent and combined contribution of physician preferences and prejudices on clinical decision making presents a formidable methodological challenge. Even the most sophisticated multivariate modeling of large patient databases is unlikely to overcome well-recognized problems of collinearity. Possibly the only way to estimate the independent and joint contribution of co-varying patient characteristics (like age, gender, race, and socioeco-

nomic status) on clinical decision making is to undertake carefully controlled randomized experiments where selected social attributes of patients are deliberately manipulated to estimate their unconfounded contribution. Such a controlled experiment is described in some detail below, and new results are presented.

This paper addresses the following two questions:

1. Do selected patient attributes (specifically the patient's age, gender, race, and socioeconomic status) independently influence a physician's clinical decision-making (specifically, the number and type of diagnoses considered, the level of certainty adhering to them, and the types of tests that would be ordered)?
2. Do characteristics of physicians alone (specifically their gender, age, race, and medical specialty), or in combination with attributes of patients, also affect medical decision-making?

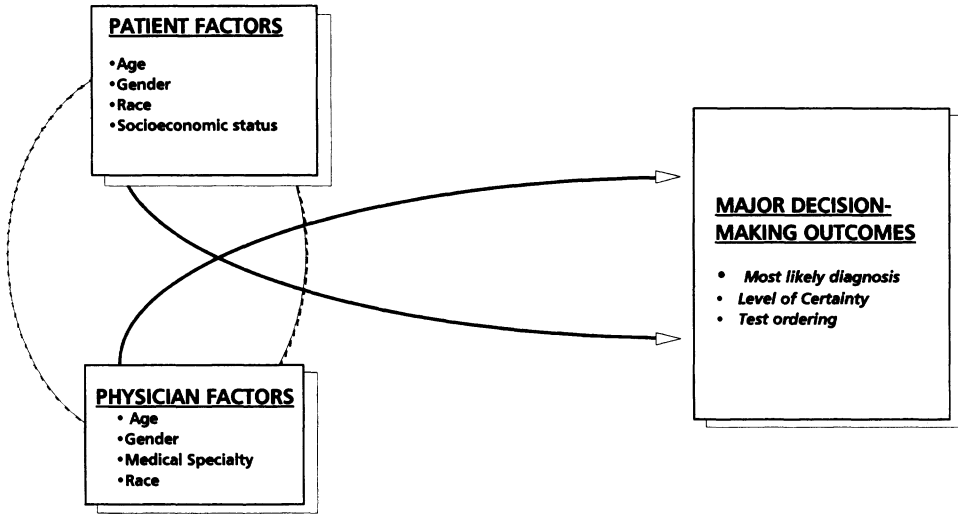
RESEARCH METHODS

Overview of Study Design

We used professional actors and actresses to realistically portray medical encounters on videotape. In one scenario the "patient" presented with polymyalgia rheumatica or PMR (a musculoskeletal condition); in the other, the patient presented with signs and symptoms indicating depression. Each scenario was taped repeatedly, systematically varying the patient's age, race, gender, and socioeconomic status. The videos were shown to a sample of 128 physicians, stratified according to race, gender, experience, and medical specialty. After viewing one selected variant of each scenario, study physicians were asked questions concerning diagnosis, treatment, referral, and medication for the patients portrayed in the video. After recording a diagnosis and intended treatment plan each physician completed a self-administered questionnaire measuring aging, sexism, and racism, and we elicited attitudes toward several aspects of medical practice.

The video variants were systematically distributed across physician strata to give us data on the greatest possible variety of combinations of patient characteristics (on video) and physician characteristics (actual). This ran-

FIGURE 1. Working Model Of Medical Decision Making



Note: The shaded area reflects the background presence of organizational influences (e.g., managed care organizations, governmental regulations, clinical guidelines, etc.) which are thought to also affect decision-making. These background influences are not the subject of this paper.

domized experimental design Cochran and Cox 1957; Hinkelman and Kempthorne 1994) allowed us to evaluate individually and simultaneously a large number of factors that may influence physicians' medical decisions, to examine interactions among factors, and to achieve optimal statistical power in a cost-effective way. More detailed description of the research approach is provided elsewhere (McKinlay, Potter, and Feldman 1996; McKinlay et al. 1997 Feldman et al. 1997.

Medical Scenarios

The videotaped scenarios portrayed a common encounter between a patient and a physician, occurring in a primary care physician's office, in which the "patient" presents symptoms of either PMR or depression. These particular conditions were selected for several reasons: (1) they represent different types of illnesses (one condition is considered "medical," while the other is more emotional or psychological); (2) they are common among the elderly and usually first presented to a primary care physician (internist or family practitioner); (3) they are usually symptomatically diffuse, suggesting a range of possible diagnoses; and (4) they also admit a range of test ordering, treat-

ment, and referral possibilities. Standardized dialogue and nonverbal behavior were based on actual cases provided by two experienced physicians. The "patient" was enacted by several different professional actors and actresses chosen to represent the characteristics of research interest (see below). The physician character was a white, middle-aged male in all variants. A third scenario, in which the patient presents with dyspnea (breathlessness), was also taped for showing between the PMR and depression scenarios in order to quench the viewer's impressions and ensure independent responses to the two scenarios of research interest. For presentation to the physician-subjects, the vignettes were sorted into 16 sets of three with complementary patient characteristics, each set running consecutively on a single videotape with PMR first, dyspnea second, and depression third. Physician subjects were not questioned in detail about the dyspnea scenario.

The final scripts for the videos were developed from transcripts of tape-recorded role playing sessions by a group of physician collaborators with considerable clinical experience in the areas of PMR and depression. Sufficiently different symptoms of these two medical conditions using lay descriptions (e.g., "lack of oomph") were carefully embedded in

the natural dialogue of the “physician” and the “patient” on the videotape (Table 1). Close attention was paid to maintaining highly standardized verbal and nonverbal behavior in each of the videos to ensure that the professional actors and actresses did not depart from their assigned roles and that they varied only on the experimentally controlled characteristics. Clinicians currently seeing patients with the conditions of interest were present during all filming sessions to ensure realism and authenticity. Other clinical colleagues who subsequently reviewed the videotapes (prior to beginning the fieldwork with the physician subjects) commented on their clinical authenticity. Indeed, during the fieldwork numerous physicians volunteered comments indicating that they believed the “patients” on the videotape were real cases and quite typical of real patients seen in their practice. The methodological importance of this emphasis on clinical authenticity is discussed below in relation to external validity.

Our analytic objective was to assess the influence of patient and physician factors for each scenario separately, rather than comparing diagnostic and treatment variables between the two scenarios. We did, however, look for consistent patterns of response in the two scenarios, as evidence of reproducibility of the results for different medical conditions. Data on the “patient” with dyspnea were not analyzed because that condition was not central to the research questions of interest.

Factorial Design

A complete factorial design includes at least one data point representing every combination of levels of all factors. In the case of two factors, the statistical model for a single data point would be

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

where *i* and *j* indicate the levels of the two factors for that data point; *y_{ij}* indicates the response; *μ* indicates the mean response; *α_i* and *β_j* indicate the independent (“main”) effects of the two factors; and *γ_{ij}* indicates the interactive effect peculiar to the combination of factor levels (*i, j*). For three or more factors, higher-order interaction terms (e.g., *δ_{ijk}*) enter the equation. A complete factorial design permits the evaluation of every main effect and every interaction. In the present experiment a complete factorial design would have required every combination of patient characteristics to be viewed by at least one physician with every combination of physician characteristics, for a total of 2⁴ × 2⁴ = 256 viewings. The supply of physicians (even in the Boston area) is simply insufficient for a complete design (see below), but with a carefully selected sample of 128 physicians to fill the required cells we did achieve a high degree of coverage of the possible combinations, allowing us to evaluate all main effects and all two-factor interactions. Higher-order interactions, many of which would not have been estimable or interpretable in any case, were not of research interest and were not included in our statistical analysis.

Patient Factors (Treatments)

Each of the four patient attributes—gender, race, age, and socioeconomic status—was dichotomized, producing altogether 2⁴ = 16 combinations. Each combination was portrayed by a different actor or actress. Male or female gender and African American or White race were represented by actors of the appropriate gender and race. Age was specified as either 67 (young-old) or 79 years (elderly) in the character synopsis at the beginning of the videotape, and the actor or actress was of that

TABLE 1. Experimental Conditions Enacted on the Videotapes

1. Polymyalgia Rheumatica (PMR):	2. Depression:
• Lacking energy • Can't get going • Stiffness upon awakening • Weight loss • Shoulder and hip girdle symptoms	• Some constipation • Appetite loss • Weight loss • Lacking “oomph” • Sleeplessness (afternoon naps) • Hopelessness

Note: The above symptoms were embedded in the dialogue between the “physician” and the “patient” in the two professionally acted videotapes (PRM or Depression) that were presented to the experimental subjects (practicing physicians).

age. The patient's medical insurance coverage was specified as Medicaid only or Medicare plus supplemental coverage in the character synopsis. The insurance information together with the patient's manner and dress formed a composite representation of socioeconomic status.

Physician Factors and Subject Recruitment

The four dichotomous physician characteristics were gender, race, medical specialty, and experience. Experience was measured by the number of years since graduation from medical school, dichotomized as 15 years or less and 16 years or more. Physicians were randomly selected from two medical specialties (family practice and internal medicine) and from two races (African American or White). The limited availability of female and African American physicians in the sampling area required us to draw a smaller sample in those two groups. The final sample included 80 men and 48 women, with 96 White and 32 African American physicians. Sampling in multiples of 16 allowed us to present each of the 16 patient characters the same number of times within strata of each physician factor.

The initial sampling frame, consisting of primary care physicians (internists and family practitioners) in the New England region, was chosen to maximize the generalizability of inferences while retaining a feasible research design. The two medical specialty areas were similar in relation to the type of medicine practiced and that often encountered the common patient complaints to be simulated in this study, but were also characterized by a high degree of variability in the major variables of interest.

In order to be eligible for the study physicians (the experimental subjects) had to: (1) be in either internal medicine or family practice (seeing patients ≥ 50 percent of the time); (2) have trained and completed a medical residency in the United States (no International Medical graduates were selected); (3) be actively providing clinical care (in contrast to research, teaching, or administration); and (4) be regularly encountering older patients. It proved surprisingly difficult, even in the physician-abundant Northeast, to locate sufficient subjects for certain design categories (e.g., older female, African American physicians)

who could then be screened for eligibility for the study. Random sampling of eligible physicians, while possible for some categories, was necessarily abandoned in favor of convenience (snowball) procedures for the "rarer" subjects (e.g., older African American internists). This was necessary to protect the internal validity of the study. Field travel as far as Washington, DC was required to obtain some eligible subjects. Calculation of an overall response rate under these circumstances could be misleading. We do not believe that mixed approach to subject recruitment for this experiment affects the validity of the study results.

A letter of introduction was sent to each eligible prospective participant physician, inviting participation in the study. The letter stated that contact would be made by telephone, and, if the prospective participant was willing, an appointment would be set up at the participant's convenience to view the video and complete the in-person interview.

Data Collection

All data collection was conducted by a trained interviewer at the physician's practice workplace during regular hours. After obtaining informed consent, the interviewer showed a set of three videotaped scenarios, chosen in advance at random from among those not yet shown to any physician with the same characteristics, and administered a structured questionnaire after each one. The order of scenarios was always PMR first, dyspnea second, and depression third. Finally, after diagnostic and treatment information was obtained, participants completed a self-administered questionnaire containing validated scales measuring agism, sexism, and racism (which we reasonably assumed could affect clinical decision making). The physician participants were not informed of the study hypotheses. The time required for viewing the videotape and completing all questionnaires was approximately one hour. The physician was offered \$100 as reimbursement to partially offset lost revenue for this time. This was provided to the physician subject upon completing the interview.

Outcome Measures

The questionnaires contained both multiple-

choice and open-ended questions designed and field tested to facilitate collection of valid and reliable responses. First, the physician subjects were asked to list all possible diagnoses and assign a probability to each diagnosis on a scale of 0 to 100 percent. Second, the physicians were asked to list all tests that he or she would likely recommend for the patient. Third, the subjects were asked to provide treatment recommendations, such as medications, referrals, or other forms of patient management. Finally, the physician completed a self-administered questionnaire including sociodemographic information and addressing several aspects of medical practice, such as attitudes toward medical testing, concern with malpractice, comfort with uncertainty, cost-consciousness, and willingness to disclose information. The self-administered instrument was pilot-tested on 20 physicians and reviewed by the five member advisory panel before use in the field.

Three major dependent variables were analyzed: the primary diagnosis, defined as the diagnosis given the highest probability; the level of certainty attached to the primary diagnosis; and number of diagnostic tests or procedures recommended by the physician subject.

Statistical Analysis

An exhaustive analysis of the effects of interest in this randomized experiment would involve a large number of hypothesis tests:

1. four patient factors;
2. four physician factors;
3. six pairwise interactions between patient factors;
4. six pairwise interactions between physician factors; and
5. sixteen pairwise interactions between patient and physician factors.

The total of 36 hypothesis tests at $p = .05$ would make the analysis vulnerable to false inference, with nearly 2 ($.05 \times 36 = 1.80$) Type I errors expected per outcome variable. To protect against excessive Type I error, we grouped the variables as listed above and conducted a hierarchical model-building analysis, adding each group only if it improved the model significantly ($p < .05$) as judged by the F statistic for continuous outcome variables or χ^2 statistic for dichotomous outcomes. Interaction groups were tested only if the main effects involved

were already included in the model. Individual independent variables were examined only in the context of a significant group. This model-building strategy reduced the expected number of Type I errors to an acceptable level per outcome variable ($< .05 \times 5 = .25$).

Measures to Enhance External Validity

Although our experimental design enhances internal validity, the use of hypothetical "patients" obviously threatens external validity (whether a physician's response to videotaped encounters reflect his or her usual behavior in everyday real practice encounters). Four precautionary steps were taken to foster external validity. First, as discussed above, considerable effort was devoted to ensuring the clinical realism of the videotape presentation. This was achieved by using professional actors and actresses and by filming with experienced clinicians always present. Second, the physicians viewed the tapes in the context of their practice day (not at a professional meeting, a course update, or even in their homes). In other words, it was likely they saw real patients before and after they viewed the "patient" in the videotape. Third, the physicians were specifically instructed at the outset to view the "patient" as one of their own patients and to respond as they would typically respond in their own practice. Fourth, physicians also were asked during the interview whether their treatment of the videotape "patient" would be the same as the treatment of one of their own patients. If differences were reported, they were asked to describe them in detail.

RESULTS

For clarity of presentation two points should be emphasized regarding the organization of our research findings. First, two separate experiments were conducted on the same sample of 128 practicing physicians. One experiment depicted the presentation of symptoms consistent with polymyalgia rheumatica (PMR)—this was termed "the PMR Video." A second experiment presented symptoms designed to be consistent with a diagnosis of depression—termed "the Depression Video." As expected, there was inevitable overlap between these two presentations. However,

sufficiently differentiated symptoms were embedded in each video to permit separation of diagnoses (Table 1). Second, the data from these two experiments must be analyzed separately because the responses, coming from the same sample of 128 physicians (the experimental subjects), are obviously correlated. No analyses combining data from these two separate experiments are presented, as their interpretation would be compromised.

Experiment 1: The Polymyalgia Rheumatica Video

Table 2 summarizes the range of most likely diagnoses given by the 128 physicians viewing the video depicting a presentation of symptoms consistent with PMR. It is immediately evident that the overwhelming diagnosis is depression (65.6 percent), and of the remaining diagnoses, none exceed 13 percent—including PMR itself which was the leading diagnosis for only 7 percent of the physician subjects, attaining only third rank in Table 2. Analyses focusing specifically on PMR as a most likely diagnosis were necessarily limited because of insufficient data. Fisher’s (exact tests were used to examine any relationship between patient and physician characteristics, and no PMR diagnosis or significant associations emerged (Agresti, A. 1992). Because of the overwhelming selection of depression as the most likely diagnosis for the PMR presentation and the small number of physicians selecting any other diagnoses, the most likely diagnosis measure was dichotomized into depression and all other likely diagnoses combined.

A relevant question for this scenario is whether the participating physicians considered a diagnosis of PMR at all (regardless of probability). Fifty-two (40.6%) of the 128

physicians mentioned PMR as a possible diagnosis. The full model did not converge because of a zero cell frequency (no black male family practitioner mentioned PMR as a diagnosis). However, from partial models that did converge, the results are consistent with those presented for a primary diagnosis of depression, as indicated below.

Table 3 summarizes the main results for this first experiment across the three key decision-making outcomes (the most likely diagnosis, the level of certainty attached to the diagnosis, and the number of tests that would be ordered). The first outcome (the most likely diagnosis) was determined from the range of eleven diagnoses offered by the physician subjects (Table 2).

Research Question 1: Do Patient Attributes Influence Physician Decision-making?

It is clear from Table 3 that the main effects and two-factor interactions of the four patient attributes included in the experiment (age, race, gender, and socioeconomic status) do not affect any of the three outcomes studied (the most likely diagnosis, physician’s level of certainty, and the number of tests likely to be ordered). For the PMR video, and contrary to expectation, patient attributes that have been epidemiologically linked to the disease do not appear to influence physician decision-making.

Research Question 2: Do Physician Characteristics alone or in combination with patient attributes affect physician decision-making?

Table 3 reveals that none of the physician characteristics studied (physician’s age, race,

TABLE 2. Experiment I (PMR Video): Distribution of Most Likely Diagnosis (N = 128, 1997–98)

Most likely diagnosis	Frequency	Percent
Depression	84	65.6
Non-specific arthritis M-S disease	17	13.3
PMR	9	7.0
Nutritional Problem	9	7.0
Non-malignant GI condition	3	2.3
Other Conditions	6	4.6
(1 physician subject reporting each)*		
TOTAL	128	100.0

* Thyroid disorder, malignancy, COPD, Chronic Fatigue Syndrome (2) Malaise

TABLE 3. Partitioned Analyses And Resulting Tests For Three Outcomes: Experiment I: PMR Video (N = 128, 1997–8)

Source	Main Outcomes											
	Primary diagnosis				Certainty				Number of tests			
	df	G ²	ΔG^2	P	SS	MS	F	P	SS	MS	F	P
Patient main effects	4	1.39	1.39	n.s.	1.81	.45	.95	n.s.	2.45	.61	1.39	n.s.
Patient 2-factor interactions	6	12.79	6.62	n.s.	4.91	.82	1.73	n.s.	4.24	.71	1.60	n.s.
Physician main effects	4	6.18	4.79	n.s.	2.50	.62	1.32	n.s.	.82	.20	.46	n.s.
Physician 2-factor interactions	6	15.25	2.46	n.s.	.97	.16	.34	n.s.	5.87	.98	2.22*	0.04*
Patient/physician 2-factor interactions	16	22.38	7.13	n.s.	10.35	.65	1.36	n.s.	7.53	.47	1.07	n.s.
Error	91				43.16	.47			40.06	.44		
TOTAL	127				63.70				60.97			

*P < 0.05

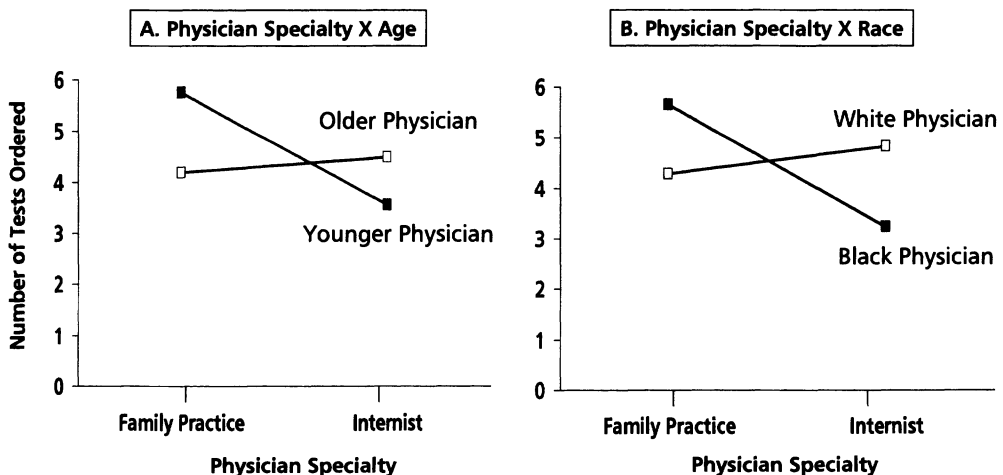
Key: df = Degrees of freedom; ΔG^2 = difference of G² by adding next set of variables; G² = -2 log likelihood (distributed as chi-square statistics); SS = sum of square; MS = mean square; P = p - value; F = f-statistics

gender, or medical specialty) had a direct (i.e., main) effect on any of the three outcomes investigated. However, for the number of tests likely to be ordered, there were significant interaction effects among these physician characteristics. Of these six interactions, the two contributing most to the significant mean square (.98) are depicted in Figure 2. Both of these interactions involve the study physician's medical specialty (whether a family practitioner or an internist). Figure 2a depicts the strong interaction between the study physicians' specialty and their age. Among the older physicians, medical specialty appears to have little effect on test ordering. In contrast, the younger physicians in family practice show markedly

higher test ordering (an average of 5.75 tests) compared to internists (an average of only 3.56 tests).

Figure 2.b depicts the much stronger interaction between physician's race and medical specialty. Among the white physicians, medical specialty had only a small effect on test ordering, with internists ordering slightly more tests than family practitioners (4.88 versus 4.28). Among the black physicians however, the trend is reversed, with a larger difference—family practitioners ordering an average of 5.66 tests and internists averaging only 3.22 tests.

There were no significant interactive effects between the physician characteristics and the

FIGURE 2. Interactive Effects between Physician Characteristics on Test Ordering, (Experiment I: PMR Video: N = 128, 1997–98)

patient attributes for any of the three main outcomes (most likely diagnosis, level of certainty, and the number of tests likely to be ordered).

A partial model (excluding physician interactions) using possible PMR diagnosis as the dependent variable did not reveal any patient effects and significant physician main effects of race and specialty only ($p < .01$, 1 df each). Family practitioners were half as likely to mention PMR compared to internists (26% vs. 55%), and African American physicians were much less likely to mention PMR than their white counterparts (19% vs. 48%).

Experiment II: The Depression Video

Table 4 summarizes the range of most likely diagnoses given by the physician subjects for the depression videotape. The most preferred diagnosis for this video, chosen by just over half of the physicians (53.9 percent), is depression. The frequency with which the other likely diagnoses were listed was too small to permit separate analysis (the presence of empty cells prevented model convergence). Moreover, depression was listed as a possible diagnosis by 121 of the 128 physicians, making analysis of this outcome also not feasible due to occurrence of empty cells. The first outcome was therefore dichotomized as depression versus all other primary diagnoses.

Research Question 3: Do patient attributes influence physician decision-making?

Table 5 summarizes the major results for the second experiment focusing on depression, across the same three outcomes. The first outcome (the most likely diagnosis) was determined, in Experiment 1, from the range of ten

most likely diagnoses offered by the physician subjects.

It is clear from Table 5 that none of the patient attributes had a direct or two-factor interactive effect on any of the three outcomes considered. In other words, also for this second experiment concerning depression, patient attributes that have been epidemiologically linked to the disease do not appear to affect physician decision-making.

Research Question 4: Do physician characteristics alone or in combination with patient attributes affect physician decision-making?

Table 5 reveals that physician characteristics only affect the first outcome directly (main effect). Specifically, white physicians are nearly twice as likely to diagnose depression as their black physician counterparts (61 percent versus 34 percent). Internists are also more likely to diagnose depression than are family practitioners (65.5 percent versus 42.8 percent).

Interactions among the physician characteristics do not have any statistically significant effect on any of the three outcomes. However, two-way interactions between the physician characteristics and patient attributes do appear to affect the most likely diagnosis. The two interactions contributing most to the mean square are depicted in Figure 3, and both involve the physician's age. Figure 3a shows that younger physicians are more likely to diagnose depression in "young-old" patients (68.5 percent versus 37.5 percent in other patients), while older physicians show an equivalent preference for a diagnosis of depression in elderly patients (65.6 percent versus 45.2 percent in "young-old" patients).

Figure 3b depicts an equivalent interaction between a physician's age and patient's gender.

TABLE 4. Experiment II (Depression Video): Distribution of Most Likely Diagnosis (N = 128, 1997-98)

Most Likely Diagnosis	Frequency	Percent
Depression	69	53.9
Non-malignant GI condition	22	17.2
Non-specific arthritis M-S disease	14	10.9
Malignancy	13	10.2
Nutritional Problem	4	3.1
Other Diagnoses	4	3.1
(one physician subjects reporting each)*		
TOTAL	128	100.0

*Mental Health disorder, undefined Dental problem, Hypertension (2)

TABLE 5. Partitioned Analyses And Resulting Tests For Three Outcomes: Experiment II: Depression Video (N = 128, 1997–8).

Source	Main Outcomes											
	Primary diagnosis				Certainty				Number of tests			
	df	G	ΔG^2	P	SS	MS	F	P	SS	MS	F	P
Patient main effects	4	2.40	2.40	n.s.	3.41	.85	1.96	n.s.	1.66	.41	1.17	n.s.
Physician 2-factor interactions	6	19.75	1.61	n.s.	.004	.0007	.0017	n.s.	3.39	.57	1.60	n.s.
Physician main effects	4	18.14	15.73	.001**	2.18	.54	1.25	n.s.	1.30	.32	.92	n.s.
Patient/physician 2-factor interactions	6	27.38	7.63	n.s.	1.88	.31	.72	n.s.	2.78	.46	1.31	n.s.
Patient/physician 2-factor interactions	16	58.54	31.16	.013*	8.37	.52	1.20	n.s.	8.20	.51	1.45	n.s.
Error	91				39.55	.43			32.19	.35		
TOTAL	127				55.39				49.52			

*P < 0.05

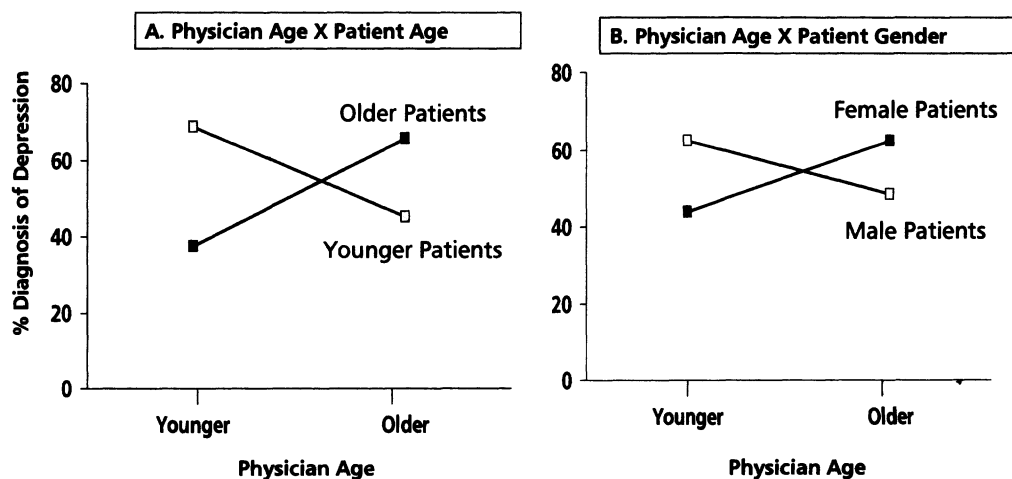
Key: = Degrees of freedom; ΔG^2 = difference of G^2 by adding next set of variables; G^2 = $-2 \log$ likelihood (distributed as chi-square statistics); SS = sum of square; MS = mean square; P = p – value; F = f-statistics

Younger physicians are more likely to diagnose depression in male patients (62.5 percent versus 43.8 percent in females), while the older physicians in the study were more likely to diagnose depression for female patients (62.5 percent versus 48.4 percent in males).

CONCLUSIONS AND DISCUSSION

Prescriptive decision theorists employ an idealized view of how physicians, as rational actors utilizing Bayesian reasoning, ought to make appropriate medical decisions. In their view patient attributes that are epidemiologically linked to the prevalence of a medical con-

dition should be reflected in medical decisions for patients with those attributes presenting with symptoms of that condition. The two medical conditions simulated in the present experiments are well described in the medical literature. Polymyalgia rheumatic (PMR) is a cause of musculoskeletal symptoms in older patients (with an estimated prevalence of 500 per 100,000 for the population over 50 years, or about 1 in 200 persons). Numerous studies report that this nonfatal disease of unknown etiology predominates in females (by a ratio of approximately 5:2) and overwhelmingly within White populations (Salvarani et al. 1995; Gran and Myklebust 1997; Labbe and Hardouin 1998). Depression is also a common affective

FIGURE 3. Interactive Effects of Physician and Patient Attributes on Most Likely Diagnosis (Depression v. Other) (Experiment II: N = 128, 1997–98)

disorder, which is reported (or commonly thought) to change with age, be more prevalent among women and among lower socioeconomic groups, and be less prevalent among African Americans (Blazer, Hughes, and George 1987; Klerman and Weissman 1989; Blazer et al. 1994; Brown 1990). Since there was no observed influence of the patient attributes studied (age, race, gender, and socioeconomic status) on any of the aspects of clinical decision-making considered (including the most likely diagnosis, the level of certainty, and the number of tests likely to be ordered) for either of the medical conditions, it does not appear that the physician subjects in this carefully controlled experiment were behaving as good Bayesians. The consistent absence of effects of patient attributes on medical decision-making across the two separate experiments is noteworthy.

In contrast to the patient attributes (which prescriptive theorists believe should have an effect), selected physician characteristics appeared to influence medical decisions (which prescriptive theorists believe should not have an effect). For experiment 1 (the PMR Video) the effects were only on the number of tests likely to be ordered and only involved an interaction between physician characteristics (physician specialty and physician race). Table 6 suggests that an increased threat of malpractice (litigaphobia) among family practitioners and African American physicians may explain these effects. However, in a partial model, main effects of physician race and specialty significantly affected any mention of a PMR diagnosis, so that malpractice concerns may provide a partial explanation at best.

With respect to experiment 2 (the Depression Video), we noted the interactive effect of physician's age and patient gender—older physicians were more likely than younger physicians to diagnose depression in female patients, and the young physicians were more likely than older physicians to do so in male patients. This may be a cohort effect of

medical training. In earlier decades it was thought that depression was commonly a female condition (it was once erroneously believed to be influenced by menopausal change). More recent evidence suggests that depression is a condition that also afflicts males, among whom it is often overlooked and underdetected (Garrard et al. 1998). More recently trained physicians may have been exposed to this new evidence and attuned to symptoms of depression in male patients.

It is difficult to explain the interaction between the physician's age and the patient's age. Social science theory suggests that individuals are more comfortable and have an affinity of interest with those who are of the same general sociodemographic category (often referred to as "status homophily"). However, there was, on average, more than 20 years separating the younger physicians from the "young-old" patients and the older physicians from the elderly patients. In each case, the age of the patients who were likely to receive a diagnosis of depression was probably close to the age of the physician's own parents.

In interpreting these somewhat unexpected findings, two constraints on the conduct of the study must be highlighted. First, for reasons of cost efficiency, and as described elsewhere, the same study physicians were used for both experiments. Even within the Boston area, with its reported surfeit of physicians, it was difficult to recruit 128 eligible physicians so as to fill all the cells required by the factorial design; a sample of 256 would have been logistically difficult and prohibitively expensive. It is possible that our sample of 128 study physicians was sufficiently atypical to yield non-generalizable results.

Second, in implementing the experimental design, the ordering of the videos was not randomized, and the experiment 1 video (PMR) always preceded the experiment 2 (Depression) video. While this in no way affects the integrity of each experiment and its results, this nonrandom ordering may partially

TABLE 6. Mean (SE) Malpractice Concern Score By Physician Specialty and Physician Age And Race (N = 124, Experiment I: PMR Video; 1997–1998)*

Physician Specialty	A. Physician's Age		B. Physician's Race	
	Younger	Older	White	Black
Family Practice	4.6 (0.36)	5.4 (0.43)	4.8 (0.34)	5.4 (0.50)
Internal Medicine	3.1 (0.29)	4.0 (0.41)	3.7 (0.29)	3.0 (0.48)

*Four Physician Subjects did not respond to a separate Self-Administered Completed Questionnaire

explain the different distributions of most likely diagnosis summarized in Tables 2 and 4. Despite this, it should be emphasized that the results of these two experiments cannot be directly compared or combined because of the correlated responses of the physician subjects.

Third, elsewhere we have identified fundamental flaws in the Bayesian approach to medical decision-making, using the widely reported gender difference in coronary heart disease (CHD) as an example (McKinlay, Potter, and Feldman 1996). Using Bayesian reasoning, the diagnosis of coronary heart disease for a particular female patient should be based on the following posterior probability (here D = patient-specific diagnostic data such as blood pressure or a test result, C is a constant, and F indicates female):

$$\Pr(\text{CHD} \mid D, F) = C \times \Pr(\text{CHD} \mid F) \Pr(D \mid \text{CHD}, F)$$

In this equation, $\Pr(\text{CHD} \mid F)$ is the true prior coronary heart disease prevalence among females, and $\Pr(D \mid \text{CHD}, F)$ is the likelihood of observing diagnostic data D for a female coronary heart disease patient. In this formulation, the prior estimate of gender-specific coronary heart disease prevalence [$\Pr(\text{CHD} \mid F)$] is assumed to be independent of the likelihood of observing patient-specific diagnostic data from a subject with CHD [$\Pr(D \mid \text{CHD}, F)$]. In reality, these two probabilities are seldom independent, as our best estimate of the prior probability (the observed rate) is based on the diagnostic process itself. This circularity in Bayesian reasoning can easily lead to underdiagnosis of disease in population groups observed to have lower disease rates (McKinlay et al. 1996).

The absence of any influence of patient attributes in the experiments reported here may indeed indicate that the study physicians are not acting as "good Bayesians." Rather, the null result reported may reflect the physicians' appropriate reliance only on specific signs and symptoms. In using these specific signs and symptoms (without weighting them in importance according to patient attributes such as age or gender), physicians may be less likely to under- or over-diagnose disease in particular patient groups. It is possible that acting as a "good Bayesian" makes for a "bad" diagnostician.

Finally, our experimental approach affords

distinct methodological advantages over the usual observational approaches taken in most studies of medical decision-making. The randomized factorial design provides an efficient, completely unconfounded comparison of each factor studied, whereas uncontrolled, non-randomized observational studies unavoidably are plagued by confounding factors (e.g., race and socioeconomic status for patient attributes, and age and gender for physician characteristics) which cannot be removed by large numbers or sophisticated analyses (Shulman et al. 1999). The factorial experimental design makes extremely efficient use of only 128 physician subjects, providing adequate statistical power for most comparisons. Most importantly, our study and other separate experiments (Freund et al. 1995; McKinlay et al. 1996; McKinlay et al. 1997; Feldman et al. 1997; Krupat et al. 1999) demonstrate the feasibility and utility of rigorous randomized experiments in the field of medical decision-making and suggests that future work on factors influencing physician behavior should move beyond observational approaches, with their inevitable limitations of confounding, equivocal results and prohibitive cost.

At least two important implications follow from the work described in this paper: (1) formal Bayesian approaches (prescriptive), while presently dominating the field of coronary heart disease, are asociological, miss much of what actually goes on during a clinical encounter, and are fundamentally flawed (the nonindependence of apriori and posterior probabilities); and (2) observational approaches, which also continue to dominate the study of the doctor-patient encounter, suffer from unavoidable confounding factors (which can never be successfully eliminated with even the most sophisticated multivariate manipulations) and should not be encouraged because a superior methodological alternative exists (factorial experimentation).

NOTE

1. While appreciating the many contributions of prescriptive theorists, we have deliberately chosen to employ a descriptive approach. This decision has consequences: For example, we omit any consideration of such issues as base rates, pre-test probabilities, sensitivity and specificity, and posterior

adjustments. These issues emanate from the prescriptive tradition and its underlying assumptions (an approach we elected not to follow) and have little relevance from the descriptive viewpoint. Moreover, this paper focuses as much on physician attributes and their influences on clinical decision making as it does on patient characteristics (which provide most of the information for assigning base rates).

REFERENCES

- Agresti, Alan. (1992). "A Survey of Exact Inference for Contingency Tables." *Statistical Science*, 7(1), 131-177.
- Balint Michael. 1957. *The Doctor, His Patient and the Illness*. New York, NY: International Universities Press.
- Bittner Egon. "Police Discretion in Emergency Apprehension of Mentally Ill persons." 1967. *Social Problems* 14:178-292.
- Blazer, Dan G., Dana C. Hughes, and Linda K. George. 1987. "The Epidemiology of Depression in an Elderly Community Population." *Gerontologist* 27:281-37.
- Blazer, Dan G., Ronald C. Kessler, Kahterine A. McGonagle, and Marvin S Swartz. 1994. "The Prevalence and Distribution of Major Depression in a National Community Sample: The National Comorbidity Study." *American Journal of Psychiatry* 151(7):979-86.
- Bloom, Samuel and Pamela Summey. 1976. "Models of the Doctor-patient Relationship: A History of the Social System Concept." P. 17 in: *The Doctor-Patient Relationship in the Changing Health Scene* edited by Eugene Gallagher. Washington, DC: U.S. Dept. of Health Education and Welfare.
- Brown, D.P. 1990. "Depression Among African Americans: An Epidemiologic Perspective." In: *Handbook of Mental Health and Mental Disorder Among Black Americans*. Riuz, Dorothy S. (Ed). New York, NY: Greenwood Press.
- Clark, John, Deborah Potter, and John McKinlay. 1991. "Bringing Social Structure Back into Clinical Decision-making." *Social Science and Medicine* 32(8):853-66.
- Cochran, William G., Gertrude M. Cox. 1957. *Experimental Designs*. New York, NY: John Wiley & Sons, Inc.
- Elstein, Arthur, Lee Shulman. 1978. *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press.
- Feldman, Henry A., John B. McKinlay, Deborah A. Potter, Karen M. Freund, Risa B. Burns, and Mark A. Moskowitz. 1997. "Non-medical Influences on Medical Decision-making: An Experimental Technique Using Videotapes, Factorial Design, and Survey Sampling." *Health Services Research* 32(3):343-65.
- Freidson, Eliot. 1962. "Dilemmas in the Doctor-Patient relationship." In Pp 207-224 *Human Behavior and Social Processes edited by AM Rose*. Boston, MA: Houghton Mifflin.
- Freidson, Eliot. 1970a. *Professional of Medicine*. New York, NY: Harper & Row.
- Freidson Eliot. 1970b. *Professional Dominance*. Chicago, IL: Aldine.
- Freund, Karen M., Risa B. Burns, Mark A. Moskowitz, Henry A. Feldman, Linda E. Kasten, Deborah A Potter, John B McKinlay. 1995. "Patient Factors Influence Physician Choices in Management of Breast Cancer." *Journal of General Internal Medicine* 10(Supplement):55.
- Garrard, J., S.J. Rolnick, N.M. Nitz, L. Luepke, J. Jackson, L.R. Fischer, C. Leibson, P.C. Bland, R. Heinrich, and LA Waller. 1998. "Clinical Detection of Depression Among Community-based Elderly People with Self-reported Symptoms of Depression." *Journal of Gerontology: Medical Sciences* 53A:M92-M101.
- Gran, Jan Tore and Geirmund Myklebust. 1997. "The Incidence of Polymyalgia Rheumatica and Temporal Arthritis in the County of Aust Agder, South Norway: a Prospective Study." *Journal of Rheumatology* 24(9):1739-43.
- Henderson, L.J. 1935. "Physician and Patient as a Social System." *New England Journal of Medicine* 212(18) 819-823.
- Kannel William B. 1995. "An Update on Coronary Risk Factors." *Medicine Clinical North American* 79(5):951-71.
- Kitsuse John I. 1968. "Societal Reactions to Deviant Behavior: Problems of Theory and Method." *Social Problems* 9:247-56.
- Klerman, Gerald L. and M.M. Weissman. 1989. "Increasing Rates of Depression." *Journal of the American Medical Association* 261:2229-35.
- Krupat, Edward, Julie Irish, Linda Kasten, Karen Freund, Risa Burns, Mark Moskowitz, and John B. McKinlay. 1999. "Patient Assertiveness and Physician Decision-making Among Older Breast Cancer Patients." *Social Science & Medicine* 49(4):449.
- Labbe, P. and P. Hardouin. 1998. "Epidemiology and Optimal Management of Polymyalgia Rheumatica." *Drugs Aging* 13(2):109-18.
- Lustad, L.B. 1968. *Introduction to Medical Decision-Making*. Springfield, IL: Thomas.
- McKinlay, John B. 1973. "On the professional regulation of change." P. 61 in *Professionalism and Social Change. Sociological Review Monographs*, 20, edited by P.J. Halmos.
- McKinlay, John B. 1996. "Some Contributions from the Social System to Gender Inequalities in

- Heart Disease." *Journal of Health and Social Behavior* 37(March):1-26.
- McKinlay, John B., Risa B. Burns, Richard Durante, Henry A. Feldman, Karen M. Freund, Brooke S. Harrow, Julie T. Irish, Linda E. Kasten, and Mark A. Moskowitz. 1997. "Patient, Physician and Presentational Influences on Clinical Decision-making for Breast Cancer: Results from a Factorial Experiment." *Journal of Evaluation in Clinical Practice* 3:1-69.
- McKinlay John B. and Lisa D. Marceau 2001. "Third Generation Studies of the Doctor-patient Relationship: Macro Determinants of Micro Processes." Presented at the annual meetings of the American Sociological Association, August 22, Anaheim, CA.
- McKinlay, John B., Deborah Potter, and Henry Feldman. 1996. "Non-Medical Influences on Medical Decision-Making." *Social Science & Medicine* 42:769-76.
- Parsons, Talcott. 1951. *The Social System*. New York, NY: Free Press.
- Pauker, Stephen G. 1982. "Prescriptive Approaches to Medical Decision-making." *American Behavioral Sciences* 9(1):502-522.
- Salvarani, Carol, Sherine E. Gabriel, W. Michael O'Fallon, and Gene G. Hunder. 1995. "Epidemiology of Polymyalgia Rheumatica in Olmsted County, Minnesota, 1970-1991." *Arthritis and Rheumatism* 38(3):369-73.
- Schulman, Kevin A., Jesse A. Berlin, William Harless, Jon F. Kerner, Shyrl Sistrunk, Bernard J. Gersh, Ross Dubé, Christopher K. Taleghani, Jennifer E. Burke, Sankey Williams, John M. Eisenberg, and Jose J. Escarce. 1999. "The Effect of Race and Sex on Physicians' Recommendations for Cardiac Catheterization." *New England Journal of Medicine* 340(8):618-26.
- Schwartz, William B., G. Anthony Gorry, Jerome P. Kassirer and Alvin Essig. 1973. "Decision Analysis and Clinical Judgment." *American Journal of Medicine* 55(3):459.
- Waitzkin, Harold and John D. Stoeckle. 1976. "Information and Control and the Micropolitics of Health Care: Summary of an On-going Research Project." *Social Science and Medicine* 10:263-76.
- Weinstein, Milton. 1980. *Clinical Decision Analysis*. Philadelphia, PA: Saunder.
- Wilson, Peter W. 1994. "Established Risk Factors and Coronary Artery Disease: The Framingham Study." *American Journal of Hypertension* 7(part 2):7S-12S.
- Zola, Irving Kenneth. "Medicine As an Institution of Social Control." 1972. *American Sociological Review* 20:478-504.

John McKinlay is Senior Vice President and Chief Scientist at the New England Research Institutes (NERI) in Watertown, Massachusetts. He is the 1995 recipient of the Leo G. Reeder Award and formerly held simultaneous Professorships in Medicine, Epidemiology and Biostatistics, and Sociology at Boston University. He is consultant to the Division of Medicine, Massachusetts General Hospital (Harvard Medical School).

Ting Lin is a biostatistician formerly with the New England Research Institutes.

Karen Freund is an Associate Professor of Medicine at Boston University and Chief of the Women's Health Unit at Boston Medical Center.

Mark Moskowitz is Professor of Medicine and Public Health at Boston University, Chief of the Section of General Internal Medicine, and Vice Chair of the Department of Medicine at Boston Medical Center.