Methods Section

Nonmedical Influences on Medical Decision Making: An Experimental Technique Using Videotapes, Factorial Design, and Survey Sampling

Henry A. Feldman, John B. McKinlay, Deborah A. Potter, Karen M. Freund, Risa B. Burns, Mark A. Moskowitz, and Linda E. Kasten

Objective. To study nonmedical influences on the doctor-patient interaction. A technique using simulated patients and "real" doctors is described.

Data Sources. A random sample of physicians, stratified on such characteristics as demographics, specialty, or experience, and selected from commercial and professional listings.

Study Design. A medical appointment is depicted on videotape by professional actors. The patient's presenting complaint (e.g., chest pain) allows a range of valid interpretation. Several alternative versions are taped, featuring the same script with patient-actors of different age, sex, race, or other characteristics. Fractional factorial design is used to select a balanced subset of patient characteristics, reducing costs without biasing the outcome.

Data Collection. Each physician is shown one version of the videotape appointment and is asked to describe how he or she would diagnose or treat such a patient.

Principal Findings. Two studies using this technique have been completed to date, one involving chest pain and dyspnea and the other involving breast cancer. The factorial design provided sufficient power, despite limited sample size, to demonstrate with statistical significance various influences of the experimental and stratification variables, including the patient's gender and age and the physician's experience. Persistent recruitment produced a high response rate, minimizing selection bias and enhancing validity.

Conclusion. These techniques permit us to determine, with a degree of control unattainable in observational studies, whether medical decisions as described by actual physicians and drawn from a demographic or professional group of interest, are influenced by a prescribed set of nonmedical factors.

Key Words. Medical decision making, simulated patients, experimental design, medical sociology

Numerous reports suggest that medical decision making—ideally a matter of symptoms, tests, and probabilities—is in fact a social transaction prone to medically extraneous influences. These nonmedical factors include personal characteristics of both patient and physician, as well as organizational characteristics of the setting where healthcare is delivered (Clark, Potter, and McKinlay 1991; Haug and Ory 1987). For example, aggressive treatment of breast cancer is reported to be less likely when the patient is over 75 years of age (Silliman et al. 1989; Chu, Diehr, Feigl, et al. 1987; Greenfield et al. 1987), despite comparable survival rates and tolerance of chemotherapy (Yancik, Ries, and Yates 1989; Early Breast Cancer Trial Collaborative Group 1988). Aggressive treatment is also reportedly less likely when the physician is older (Belanger, Moore, and Tannock 1991).

Demonstrating such influences objectively is a difficult matter, but an important one for the field of medical decision making. This is particularly true in areas of medicine where standard practice is in flux because of new biomedical or technical developments. With some diseases, such as AIDS or breast cancer, treatment choices have acquired a quasi-political flavor, owing to the emergence of an active patient advocacy movement. The rapid evolution of new organizational contexts for medical care raises additional questions concerning the incentives, constraints, and barriers that actually operate in the doctor-patient transaction.

The difficulties of doing useful research in this area can be summarized under two general headings: (1) the multiplicity of variables and (2) the shortcomings of observational studies.

The first problem arises simply because many nonmedical influences demand attention, injecting a large and potentially unmanageable set of

---

Supported by National Institutes of Health grants HL-37762 and AG-11352.

Henry A. Feldman, Ph.D. is Principal Research Scientist at the New England Research Institutes, Watertown, MA; John B. McKinlay, Ph.D. is Senior Vice President and Director, the New England Research Institutes; Deborah A. Potter, M.A. is Research Scientist, New England Research Institutes; Karen M. Freund, M.D., M.P.H. is Associate Professor of Medicine, Section of General Internal Medicine, Boston University School of Medicine; Risa B. Burns, M.D., M.P.H. is Assistant Professor of Medicine, Section of General Internal Medicine, Boston University School of Medicine; Mark A. Moskowitz, M.D. is Professor of Medicine, Section of General Internal Medicine, Boston University School of Medicine; and Linda E. Kasten, M.A. is Associate Research Scientist, New England Research Institutes. Address all correspondence and requests for reprints to Dr. Feldman at New England Research Institutes, 9 Galen Street, Watertown, MA 02172, Tel. 617/923-7747, ext. 314; FAX 617/926-8246; e-mail henryf@neri.org. This article, submitted to *Health Services Research* on January 18, 1996, was revised and accepted for publication on August 28, 1996.

independent variables into any analysis plan. Moreover, many of these variables are vague in nature. On the patient's side, these include socioeconomic status, insurance coverage, physical attractiveness, assertiveness, and medical knowledge. On the physician's side, the factors include experience; specialty; practice setting; and personal attitudes, beliefs, or concerns. On both sides age, sex, and race come into play, as well as numerous intangible factors and forms of nonverbal behavior. Scoring any but a few of these variables meaningfully in an observational study, or controlling them systematically in an experimental study, is certain to present serious problems. Gathering enough cases to analyze the independent and joint influences of such a large ensemble of predictors with adequate statistical power is likewise very difficult. Nevertheless, the omission of any of these factors can stand out as a glaring oversimplification in any particular area of medical decision making.

The second major impediment to definitive research in nonmedical influences is that most of the work in this field has necessarily been observational rather than experimental. Short of a full-scale clinical trial, it is simply not practical or ethical to intervene in the presentation of patients to physicians or in the formation of diagnostic and treatment decisions. Abstracts of *faits accomplis* are therefore the predominant source of data on medical decision making, at least where actual patients are involved. In the field of breast cancer, for example, Chu et al. (1987) drew data from community hospitals, while Ayanian et al. (1993) combined data from the New Jersey State Cancer registry, state-mandated hospital discharge abstracts, and the U.S. Census. Similarly Samet et al. (1986) analyzed the New Mexico tumor registry, and Yancik, Ries, and Yates (1989) utilized the Surveillance, Epidemiology, and End Results (SEER) database of the National Cancer Institute. In large observational studies, the physicians' decisions cannot be examined in detail. In smaller studies, even though the circumstances of a medical decision can be more thoroughly recorded, those circumstances are not under the researchers' control. Any characterizations of patient, physician, and setting are therefore necessarily approximate. Whether large or small, all observational studies are subject to the possibility of confounding, whereby such variables as race and socioeconomic status–no matter how well defined and recorded–tend to be strongly correlated and therefore largely inseparable in interpreting the results.

In this report we describe a set of techniques designed to overcome the problems just discussed.

1. A medical appointment is depicted on videotape by professional actors. The patient's presenting complaints are relatively specific (e.g., chest pain, breast lump, dyspnea, depression) but ambiguous enough to allow a range of valid interpretation.
2. The scenario is videotaped in several alternative versions, featuring patient-actors of different age, sex, race, or other characteristics. The script does not vary except for minor modifications of language appropriate to the character of the patient.
3. A random sample of actual physicians is selected to view the scenario. The sample is stratified on such characteristics as demographics, specialty, or experience. After viewing one version of the tape during office hours, the physician is asked to describe how he or she would diagnose or treat such a patient.
4. Exhaustive sampling efforts and persistent recruitment techniques are employed to achieve a high response rate, minimizing selection bias and enhancing validity.
5. Fractional factorial design is used to balance the experiment while minimizing cost. Typically, $k = 5$–8 patient factors are of interest. To videotape all 2^k possible combinations would be prohibitively expensive. Instead, a balanced subset is produced, reducing costs without biasing the outcome. The number of physician factors is typically smaller, $k = 2$–3, so that the strata are large and every video can be shown at least once in each stratum.

Taken together, these techniques permit us to present a standardized stimulus to a stratified random sample of physician subjects. Thus we can determine, with a degree of control and objectivity not possible in observational studies, how medical decisions as described by actual physicians are influenced by a substantial number of nonmedical factors.

Two major studies have been completed using this design, addressing diagnosis and treatment of cardiopulmonary complaints in one case and breast cancer in the other. Substantive outcomes of the two studies are presented elsewhere (McKinlay, Potter, and Feldman 1996; Freund, Burns, Moskowitz, et al. 1995; Burns, Freund, Moskowitz, et al. submitted; McKinlay, Burns, Durante, et al. 1997; Kasten, McKinlay, Freund, et al. submitted). In this report the methods and methodological results will be detailed, emphasizing (1) the experimental design and statistical power of the breast cancer study and (2) techniques used to enhance the response rate of physicians in the cardiopulmonary study.

METHODS

EXPERIMENTAL DESIGN

The key feature of our technique is the presentation of simulated patients on videotape to a community-based random sample of physicians. To illustrate the technique, the recently completed breast cancer study will be described in detail (Freund, Burns, Moskowitz, et al. 1995; Burns, Freund, Moskowitz, et al. submitted; McKinlay, Burns, Durante, et al. 1997; Kasten, McKinlay, Freund, et al. submitted). Diagnosis and treatment of breast cancer is an area with considerable latitude for physicians' preferences in testing, management, choice of therapy, and accommodation of the patients' wishes, and it therefore is well suited to examination by this method.

Medical Scenarios. To evoke a range of responses and to assess their determinants, we developed two video scenarios for the physician's consideration (Table 1).

In the first scenario ("prediagnosis"), the patient has been referred to a consulting physician by her primary care physician. The patient has discovered a breast "lump," but the referring physician's letter leaves it unclear whether this is a discrete mass or an area of nodularity or thickening. The clinical evidence is equivocal. The mammogram report and film (both provided to the viewing physician) show no abnormalities. This case thus maximizes the consultant's uncertainty about whether the lump is cancer and allows room for the viewer to exercise personal preferences that express whatever nonmedical influences are operating.

Table 1: Scenarios Enacted on Videotape

Scenario	*Data*	*Medical Issues*
Prediagnosis	Patient referred from primary care Possible breast lump felt by patient Questionable whether discrete mass No abnormalities on mammogram Normal physical exam	*Is it cancer?* *How to manage?*
Postdiagnosis	Patient referred for second opinion after biopsy Biopsy proved 0.8-cm infiltrating ductal carcinoma Clean margins, equivocal hormone receptors (If staging requested by viewer): Two of 29 tested nodes positive for tumor All metastatic evaluation negative	*Further metastatic evaluation?* *Axillary node dissection?* *Chemotherapy?* *Tamoxifen?* *Mastectomy?* *Lumpectomy with radiation?* *Reconstructive surgery?*

In the second videotaped scenario ("postdiagnosis"), the consulting physician sees a different patient, who has been referred for a second opinion after a positive biopsy that has established the presence of a 0.8-cm infiltrating carcinoma, with clean margins on the specimen and equivocal hormone receptors. Staging information, if requested by the viewer, indicates that 2 of 29 tested nodes are positive for tumor, and all metastatic evaluation is negative. The patient thus has Stage IIA disease, the area of least consensus among practitioners with regard to the appropriate choice of primary therapy or the need for adjuvant therapy.

The script for each scenario was developed by the three clinical investigators from Boston University Medical Center (BUMC) and was based on cases provided by two experienced clinicians. The dialogue was reviewed for authenticity by a panel of practicing physicians and edited for continuity by a professional video writer-producer. The scenes were rehearsed and performed by professional actors and videotaped under professional direction in an actual hospital office. Strict quality control procedures were followed during taping to ensure that the script was followed faithfully and that each actor playing the patient maintained her assigned characteristics, including both verbal and nonverbal behavior. Variant sequences within each scenario (see further on) were executed in a standardized fashion for later editing into the tape. One male actor played the consulting doctor in all variant scenarios and remained largely off-camera.

Patient Characteristics. The patient was portrayed by eight different female actors, each presenting a certain combination of age, race, and socioeconomic status (SES). Age was set at either 65 or 80 years, as listed in a character synopsis at the beginning of the videotape and enacted by an actor of approximately that age. Race was represented by casting an

Table 2: Patients Portrayed on Videotape: Half-Factorial Design

*Actor**	*Age*	*Race*	*SES*	*Scenarios*	*Variants*
Alice	65	White	Upper	Pre, Post	A
Annie	65	White	Lower	Pre, Post	B
Bess	80	White	Upper	Pre, Post	B
Blanche	80	White	Lower	Pre, Post	A
Edith	65	Black	Upper	Pre, Post	B
Henrietta	65	Black	Lower	Pre, Post	A
Lenore	80	Black	Upper	Pre, Post	A
Reva	80	Black	Lower	Pre, Post	B

* Pseudonym.

actor from the appropriate group. SES was expressed visually in the style of dress and verbally by minor grammatical variants of the script. SES was also implied by the specification of insurance coverage (either Medicaid or Medex/Medicare) in text form at the beginning of the tape. With all three characteristics thus dichotomized, the possible combinations numbered $2^3 = 8$, and each combination was assigned to a different actor-patient. Age-appropriate pseudonyms were assigned to the patients for convenience in presentation (Table 2).

Each actor recorded both the pre- and the postdiagnosis scenario. Additionally, each actor produced four variant video segments to be edited later into different versions of the two scenes. The variants added three more characteristics of potential importance: physical mobility, medical condition, and assertiveness (Table 3).

Physical mobility was dichotomized as the presence or absence of severe osteoarthritis of the knees, noted in the synopsis and dramatized by the actor's use of a walker as she entered the office. Medical condition was dichotomized as either absence of "other medical problems," noted in the synopsis, or presence of comorbidity in the form of combined hypertension and diabetes, noted in the synopsis and mentioned in the dialogue. Assertiveness was dichotomized as the presence or absence of a special line of emphatic dialogue near the end of the interview; the prediagnosis patient asked for full and timely information about her diagnostic evaluation ("I don't want to be kept in the dark"), while the postdiagnosis patient asked to know all her options and stated that she was "willing to do whatever is needed" to treat her cancer.

The three additional characteristics were capable of $2^3 = 8$ combinations, of which each actor enacted half, indicated as Set A or Set B in Tables 2 and 3. Assignment of Set A or Set B was made to produce a set of 32

Table 3: Character Variants Created by Video Editing

Set	*Number*	*Physical Condition*	*Medical Condition*	*Assertive Request*
A	1	Frail	Healthy	Yes
	2	Frail	Healthy	No
	3	Agile	Comorbidity*	Yes
	4	Agile	Comorbidity	No
B	1	Frail	Comorbidity	Yes
	2	Frail	Comorbidity	No
	3	Agile	Healthy	Yes
	4	Agile	Healthy	No

* Controlled hypertension and diabetes.

"characters" (out of the total of $2^6 = 64$ possibilities) balanced according to the principles of fractional factorial design (Cochran and Cox 1957; Kirk 1982). Each characteristic was represented in half of the videos, and each combination of one or two characteristics appeared exactly half of the time with each other characteristic. For example, half of the patients were 65 years old, and half were 80; of the 65-year-olds, half were white and half were black; of the white 65-year-olds, half were upper-SES and half were lower-SES, etc. This fractional design, simply by reducing the volume of production effort, permitted significant savings in the cost and burden of enacting, editing, reproducing, storing, transporting, and showing the videos.

Each of the 32 characters was portrayed in both the prediagnosis and the postdiagnosis scenario. A total of 64 videotapes was thus produced. These were assembled into 32 pairs for use in the field, each pair consisting of one pre- and one postdiagnosis scenario (Table 4). The patients in each pair always differed in age, race, SES, and assertiveness. In half of the pairs the patients differed in mobility but were alike in general health; in the other half, they differed in general health but were alike in mobility. This minor stricture was imposed by the half-replicate design and was balanced across all the other characteristics. The pairing of characters was maintained throughout the experiment.

Subjects. Physician subjects were recruited to fill four equal-sized strata defined by specialty (surgeons versus nonsurgeons) and experience (≤15 or

Table 4: Complementary Pairing of Scenarios for Viewing by Physician Subjects

	Prediagnosis Scenario		*Postdiagnosis Scenario*	
Pair	*Actor*	*Variants*	*Actor*	*Variants*
1	Alice	1	Reva	4
2	"	2	"	1
3	"	3	"	2
4	"	4	"	3
5–8	Annie	1,2,3,4	Lenore	4,1,2,3
9–12	Bess	1,2,3,4	Henrietta	2,3,4,1
13–16	Blanche	1,2,3,4	Edith	2,3,4,1
17–20	Edith	1,2,3,4	Blanche	2,3,4,1
21–24	Henrietta	1,2,3,4	Bess	2,3,4,1
25–28	Lenore	1,2,3,4	Annie	4,1,2,3
29–32	Reva	1,2,3,4	Alice	4,1,2,3

Note: The 32 pairs were assigned randomly to the 32 physicians in each stratum.

Table 5: Stratified Sample of Massachusetts Physicians Selected to View Video Scenarios

Specialty	*Eligibility (Past 5 yr)*	*Experience (Since Medical School)*	*Sample Size*
Surgery, gynecology	Performed breast biopsy, mastectomy	More than 15 yr	32
Surgery, gynecology	Performed breast biopsy, mastectomy	15 yr or less	32
Medical, radiation oncology	Cared for women with breast cancer	More than 15 yr	32
Medical, radiation oncology	Cared for women with breast cancer	15 yr or less	32
Total			128

Note: All white male, trained in the United States. Massachusetts lacked sufficient female and minority physicians to fill all strata of specialty and experience.

>15 years since graduation from medical school). The total sample was 128 physicians, or 32 in each stratum (Table 5). The 32 videotape pairs were assigned in random order to the 32 physicians in each stratum.

Subjects were selected from physicians practicing in Massachusetts with a listed specialty of surgery, gynecology, or medical or radiation oncology. Surgeons and gynecologists were eligible if they had performed both an open breast biopsy and a mastectomy in the past five years. Oncologists were eligible if they had cared for women with breast cancer in the past five years. Physicians trained outside the United States were excluded. These criteria were deliberately made liberal so that the study would be generalizable beyond those physicians with a special interest in breast cases. The sample was restricted to white male physicians because not enough female or minority physicians were available in Massachusetts in the appropriate specialties to provide adequate sample size for stratification according to those characteristics.

Physicians were selected randomly from a listing of licensed physicians provided by Business Mailers, Inc. ($n = 1{,}686$) and the Massachusetts Board of Registration ($n = 168$ in scarcer categories). Once selected, the physicians were sent an introductory letter, followed by a telephone recruitment call. Physicians were paid $100 to participate and signed an informed consent assuring them of confidentiality.

Pilot. For pilot testing, a sampling frame of 63 physicians was assembled from the geographic area of BUMC. These subjects, although unaware of the study and its purpose, were ineligible for the main study on grounds of the ultimate possibility of contamination. Their use therefore did not intrude on the sampling frame of the main study. Twenty pilot interviews

were conducted, and the questionnaire was refined for field use on the basis of the results.

Protocol. Physician subjects for the main study were selected in a sequence of random clusters, using the following procedures. A random zip code was drawn from the unused subjects' addresses. The study staff then picked a batch size, ranging from 10 to 50, depending on the current backlog of recruitment and interviewing. The newly chosen physicians were sent introductory letters, co-signed by the investigators from BUMC and New England Research Institutes, Inc. (NERI). The letter described the study without divulging its hypotheses, encouraged the physicians to participate, and notified them that a NERI staff person would be telephoning within a week to enroll them in the study. On enrollment, the physician was assigned to the next pair of videos on the list (Table 4), which had been sorted randomly for each stratum.

Interviews. To standardize the setting and to place the subject in his usual context for decision making, the interview was conducted in the physician's office during normal clinic hours. The format was a semi-structured interview, conducted by senior NERI staff with prior experience interviewing physicians. One interviewer was male and the other female; half of the physicians and half of the videotapes were assigned to each. The interviews were conducted between August 1993 and June 1994.

After watching each scenario, the physician was invited to order further diagnostic evaluation. Requests for a specific test were answered with a simulated laboratory report. The physician was allowed to act on the results by ordering further tests. After receiving test results, the physician was asked what recommendation for evaluation and follow-up he would make and what information on alternatives he would offer to the patient.

End Points. Because the results of the breast cancer study are not presented in this report, the outcomes will be sketched very briefly; full details can be found elsewhere (Freund, Burns, Moskowitz, et al. 1995; Burns, Freund, Moskowitz, et al. submitted; McKinlay, Burns, Durante, et al. 1997; Kasten, McKinlay, Freund, et al. submitted). For the prediagnosis scenario, interest centered on the physician's estimate of the likelihood of breast cancer and his diagnostic strategy. An important issue was whether tissue analysis was planned. For the postdiagnosis scenario, the principal issues were the extent of staging evaluation (axillary node dissection, metastatic evaluation), alternatives for primary therapy (breast-conserving surgery, lumpectomy with radiation, mastectomy with or without reconstructive surgery), and options for adjuvant therapy (tamoxifen, other chemotherapy, or no further treatment).

Table 6: Precision of Prevalence Estimates from Breast Cancer Study

Response Probability	*Standard Error**
.100	.027
.250	.038
.400	.043
.500	.044
.600	.043
.750	.038
.900	.027

* Based on binomial distribution, $n = 128$ physicians.

A short written questionnaire, self-administered at the end of the interview, asked for the physician's customary sources of medical information, elicited his opinion on the age and racial distribution of breast disease, and measured certain personality traits (dominance and risk taking) and attitudes toward women, the elderly, and persons of different race.

STATISTICAL ANALYSIS

The straightforward tabulation of percentages was the primary descriptive tool for summarizing the various responses of physicians to the two scenarios. These percentages had intrinsic interest, even apart from their variation among experimental groups. The precision attached to such percentages, given the sample size of 128, is detailed in Table 6.

The experiment had a factorial structure with eight main effects (six patient characteristics and two physician characteristics). Because the independent variables were all dichotomous and the end points were either dichotomous or polytomous, the appropriate inferential strategy for this study was discrete multivariate analysis, or the analysis of multi-dimensional contingency tables (Bishop, Fienberg, and Holland 1975), of which the best-suited variant was multiple logistic regression (MLR) (Hosmer and Lemeshow 1989). In most cases the statistical model was limited to the eight main effects; two-factor interactions and continuous covariates were evaluated only when needed to test a specific hypothesis raised by the investigators. Goodness-of-fit was determined by the residual chi-squared statistic (Hosmer and Lemeshow 1989). If the full model fitted poorly ($p < .05$), the endpoint was not analyzed further.

The magnitude of each significant effect was reported as a conditional odds ratio. An odds ratio of unity represents no effect. Conditional odds are defined by

$$\text{Odds [Outcome | Condition A]} = \frac{\text{Prob [Outcome occurs | Condition A]}}{\text{Prob [Outcome does not occur | Condition A]}}. \quad (1)$$

For a dichotomous main effect comprising two conditions A and B, the odds ratio is defined as

$$\text{OR [A : B]} = \frac{\text{Odds [Outcome | Condition A]}}{\text{Odds [Outcome | Condition B]}}. \quad (2)$$

For a trichotomous or higher-order polytomous factor, the main-effect odds were reported relative to overall odds:

$$\text{OR [A : All]} = \frac{\text{Odds [Outcome | Condition A]}}{\text{Odds [Outcome | Condition A or B or C or . . .]}}. \quad (3)$$

SAS software was used for logistic regression analysis (SAS Institute Inc. 1988). Odds ratios were obtained by exponentiating appropriate coefficients in the fitted model. An asymptotic 95 percent confidence interval for the odds ratio was constructed by exponentiating limits of ± 1.96 standard error about the coefficient. The hypothesis of null effect was tested by taking the difference between the odds ratio and unity (its null hypothesis value), dividing by the standard error to produce an asymptotic Gaussian deviate (z-score), and comparing the result to the normal distribution. Equivalently, the Wald statistic was computed by squaring z and comparing the result to the chi-squared distribution with 1 df (Hosmer and Lemeshow 1989).

Multiple comparisons are a concern in any study with so many dependent and independent variables. Some authors recommend formal adjustment of p-values to control the Type I error rate (Glantz and Slinker 1990). Such adjustments, however, require that some denominator be identified for the Type I error rate such as the number of end points, tables, manuscripts, or experiments under consideration. Any such choice is inevitably arbitrary and difficult to apply consistently when the study produces several related bodies of analysis in different reports. For this study the most workable denominator was the number of comparisons. Accordingly, we report the numerical p-value for each predictor and allow the reader to take responsibility for deciding what a convincing p-value is, depending on the reader's own scope of interest. In the breast cancer study, for example, if the reader uses $p = .05$ as a critical level for testing each factor in the main-effects model, the probability is $1 - (0.95)^8 = 34$ percent that any given dependent variable will show a spurious influence of at least one factor. A more stringent critical

level of $p = .01$ reduces the Type I error rate to $1 - (0.99)^8 = 8$ percent per dependent variable. To achieve a rate below 5 percent per dependent variable, it is necessary to use a critical level of $p = .006$ or smaller to test each factor.

RESULTS AND DISCUSSION

Methodological results will be emphasized here because the substantive results of the breast cancer experiment and the cardiopulmonary study can be found in other publications (McKinlay, Potter, and Feldman 1996; Freund, Burns, Moskowitz, et al. 1995; Burns, Freund, Moskowitz, et al. submitted; McKinlay, Burns, Durante, et al. 1997; Kasten, McKinlay, Freund, et al. submitted). First, we report the results of a simulation-based assessment of the statistical power of the fractional factorial design. Second, we cite some data from the breast cancer experiment that support the content and construct validity of the experimental technique. Finally, we discuss in detail the field techniques that enabled us to achieve a high response rate in the cardiopulmonary study, thus minimizing bias and enhancing validity.

PRECISION AND POWER

The breast cancer videos were shown to a sample of 128 viewers. The precision of simple descriptive percentages, given this sample size and based on the binomial distribution, is displayed in Table 6. It is evident from the tabulated standard errors that precision was best in absolute terms when the prevalence of the response was close to zero or 100 percent, and worst when the prevalence was close to 50 percent for each prong of the dichotomy.

To estimate the inferential power of the fractional factorial design, we conducted Monte Carlo simulations as follows. A data set consisting of 128 observations was constructed, with independent variables assigned according to the design (Tables 2 and 3). A binary response was attributed to each of the 128 subjects by selection of random numbers uniformly distributed between 0 and 1. The probability of response was programmed to differ between two subgroups of the 128 subjects according to a predetermined criterion; for example, in one simulation the probability of a biopsy recommendation was 50 percent in cases where the patient failed to request aggressive treatment but 75 percent when the request was made. The synthetic data were subjected to main-effects MLR, as described earlier, and tested for a statistically significant effect ($p < .05$) of the variable that had been used as a differentiating criterion

in the simulation. This process of generating and analyzing 128 observations was repeated 25 times. The fraction of repetitions in which the criterion variable was declared significant in MLR was taken as an estimate of the power of the design to detect an effect of the nature and magnitude specified in the programmed criterion.

The Monte Carlo simulation was conducted for a representative variable from each group of formally equivalent factors in the design: a fully balanced patient factor (assertiveness), a partially balanced patient factor (physical condition), and a physician factor (experience). Power was calculated in each case for a variety of baseline probabilities and effect magnitudes (50 percent versus 75 percent, 80 percent versus 90 percent, etc.). The results are displayed in Table 7.

Table 7 shows that power was excellent as long as the prevalence of the binary response differed by at least 25 percent between the groups being compared. For example, if the effect of the patient's assertiveness were to raise from .50 to .75 the physician's likelihood of recommending biopsy, the estimated power of the experiment to detect that effect would be 92 percent. Power was near-perfect (100 percent) for comparing underlying rates any farther apart. The power estimates for effects of assertiveness also apply to effects of age, race, or SES because of the symmetry of the design. Only slightly lower power, attributable to the fractional design, was estimated for effects of the patient's physical and medical condition and the physicians' strata. Again, the power was sufficient (80 percent or greater) as long as the difference in underlying rates was at least .25.

VALIDITY

We took a number of steps to ensure that the physician subjects would respond to the cases on videotape as they do to their own patients. The scripts and videotapes were reviewed for authenticity by a panel of practicing oncologists and surgeons, and modifications were made in response to their critique. With few exceptions, the physician subjects viewed the videotapes in their own offices, in the context of a practice day, rather than at home or at a professional, educational, or scientific meeting. The physicians were instructed to view the patient on the videotape as one of their own cases and to respond as they would respond in their own practice. They often made comments like, "I have a case like that," or "I saw this case this morning." The interview following each videotape also took place in the office, where the physicians normally conducted clinical practice. Each subject was asked whether the suggested management of the hypothetical patient was the same as would be provided for the physician's own patients; any differences were recorded.

Table 7: Power of Experimental Design for Breast Cancer Study

*Differentiating Variables**	*Response Probabilities*				
	Group 1	*Group 2*	*Difference*	*Power (%)†*	*95% C.I. (%)‡*
Assertiveness	.50/.50	.60/.40	.10	36	18–57
Age	.60/.40	.75/.25	.15	44	24–65
Race	.75/.25	.90/.10	.15	56	35–76
SES	.40/.60	.60/.40	.20	56	35–76
	.50/.50	.75/.25	.25	92	74–99
	.60/.40	.90/.10	.30	100	86–100
	.40/.60	.75/.25	.35	100	86–100
	.50/.50	.90/.10	.40	96	80–100
	.25/.75	.75/.25	.50	100	86–100
	.40/.60	.90/.10	.50	100	86–100
	.25/.75	.90/.10	.65	100	86–100
	.10/.90	.90/.10	.80	100	86–100
Patient's physical condition; patient's medical condition	.50/.50	.60/.40	.10	36	18–57
	.60/.40	.75/.25	.15	48	28–69
	.75/.25	.90/.10	.15	68	47–85
	.40/.60	.60/.40	.20	60	39–79
	.50/.50	.75/.25	.25	80	59–93
	.60/.40	.90/.10	.30	100	86–100
	.40/.60	.75/.25	.35	100	86–100
	.50/.50	.90/.10	.40	100	86–100
	.25/.75	.75/.25	.50	100	86–100
	.40/.60	.90/.10	.50	100	86–100
	.25/.75	.90/.10	.65	100	86–100
	.10/.90	.90/.10	.80	100	86–100
Physician's experience; physician's specialty	.50/.50	.60/.40	.10	28	12–49
	.60/.40	.75/.25	.15	28	12–49
	.75/.25	.90/.10	.15	60	39–79
	.40/.60	.60/.40	.20	68	47–85
	.50/.50	.75/.25	.25	84	64–95
	.60/.40	.90/.10	.30	100	86–100
	.40/.60	.75/.25	.35	100	86–100
	.50/.50	.90/.10	.40	100	86–100
	.25/.75	.75/.25	.50	100	86–100
	.40/.60	.90/.10	.50	100	86–100
	.25/.75	.90/.10	.65	100	86–100
	.10/.90	.90/.10	.80	100	86–100

* Variables in each group have equivalent power.

† Monte Carlo estimate from main-effects model with $p = .05$ as critical level.

‡ Confidence interval from exact binomial limits, $n = 25$.

Data from the breast cancer experiment suggest that these measures were successful in producing a valid response. Fewer than 4 percent of the physician subjects found any aspect of the patient or case presentation atypical

in either the pre- or the postdiagnosis scenario. Asked whether the patient was typical of those seen in their ordinary practice, 66 percent of the physicians responded positively for the prediagnosis scenario and 88 percent for the postdiagnosis scenario. Much of the negative response came from medical and radiation oncologists, who stated that they rarely saw patients prior to diagnosis of breast cancer, rather than finding any fault in the portrayal (McFinlay, Burns, Durante, et al. 1997).

The clinical judgments made by the physicians suggested that they were responding to the video patient in keeping with their specialty and customary case profile. For example, when asked to estimate the likelihood that the patient's breast mass was cancer, the surgeons, who often see nonmalignant breast masses, estimated a lower probability on average than the medical and radiation oncologists, who are usually consulted only after the diagnosis of breast cancer has been made. Forty percent of the surgeons recommended reconstructive surgery following mastectomy, compared with 16 percent of nonsurgical oncologists, a statistically significant difference. The physicians also responded appropriately to differences in the video patients' age, estimating the probability of five-year or ten-year survival to be substantially lower for the 80-year-old patient than for the 65-year-old (Burns, Freund, Moskowitz, et al. submitted).

Taken together, these data suggest that the physician subjects gave clinically valid answers to the questions put to them, and that the variations in clinical decision making identified by our factorial experiment can be interpreted as generalizable differences in how physicians in the sampled population care for patients.

RECRUITMENT

The cardiopulmonary study, described elsewhere in detail (McKinlay, Potter, and Feldman 1996), will serve to illustrate the intensive recruitment techniques that were employed in order to minimize bias and enhance validity. The study design called for a sample of 192 Boston area internists in active general practice, stratified and balanced by experience and practice setting (hospital, office, HMO). As in the breast cancer study, experience was divided at 15 years since medical school graduation, and the sample was restricted to U.S.-trained white male physicians because females and minorities in the appropriate categories were not available in adequate numbers in the Boston area. To recruit 192 respondents in the six strata, 210 eligible physicians in the sampling frame had to be approached, for a response rate of 91.4 percent. Only 18 eligible contacts (8.6 percent) refused. This high response rate, which

surpassed other documented attempts to recruit internists in general practice, was the result of efficient strategies for

- identifying, tracking, and confirming eligibility of subjects;
- getting past "gatekeepers" of several types to gain access to the physicians; and
- enlisting the physicians' cooperation after direct contact had been achieved.

These elements of the sampling procedure were critical in obtaining a high response rate, thus minimizing selection bias and ensuring the validity of the study. Each component will be described in detail.

Identification and Tracking

Exhaustive pursuit of potential subjects and strict enforcement of eligibility criteria contributed to the high participation rate ultimately attained in eligible subjects. Sampling began in 1987. The initial sampling frame was based on the current Physician Masterfile of the American Medical Association, which is updated every four years through a complete census of U.S. physicians, including AMA members and nonmembers, with a high response rate (85 percent) (Carter, Robyn, and Singer 1983). All Boston area internists who had completed medical residency and who practiced direct care were drawn from the Masterfile ($n = 530$). This frame proved inadequate for the design because only 85 physicians (16 percent) were classified as hospital-based, while all the rest (84 percent) were classified as office-based and none as HMO-based. The frame was therefore supplemented by complete listings of physicians obtained directly from Boston area hospitals and HMOs ($n = 481$). The total frame thus comprised 1,011 physicians.

Letters of invitation were mailed weekly to randomly selected groups of 20–40 potential respondents, who were then contacted by telephone and screened for eligibility. Telephone numbers were available for only 30 percent of the Masterfile and for none of the supplementary list. Preliminary field work indicated that the availability of phone numbers and the accuracy of addresses would be enhanced by cross-checking information with several standard sources, including *Folio's Medical Directory of Massachusetts,* the local telephone company, and the U.S. Postal Service. Business addresses rather than home addresses were used whenever possible in order to stress the professional aspect of this survey (Sudman 1985). Address changes or forwarding orders were obtained for 55 physicians (5 percent) in the sample frame. Of these,

13 were eligible and were enrolled in the study, constituting 7 percent of the final sample. By the end of data collection, only 5 out of 1,011 potential respondents remained untraceable.

Disposition of the sampling frame is displayed in Table 8. Of 1,011 potential respondents, 796 were declared ineligible, most because they were specialists or because their practice-setting stratum was already filled. Other reasons for ineligibility are detailed in Table 8. Of the 210 eligible subjects contacted, only 18 (8.6 percent) declined to participate. The rate of refusal was not significantly different between the AMA Masterfile and the supplemental list. Of those who refused, 9 (50 percent) were in office-based practice, and 15 (83 percent) were in the "more experienced" stratum.

Access

The protocol for the cardiopulmonary study specified direct screening of physicians. This requirement was potentially a serious obstacle, considering that mail and telephone messages reach less than 1 percent of management without the intervention of a "gatekeeper" (Keuch 1987; Arndt 1986). The research staff found that larger organizations distinguished between requests for medical versus nonmedical information. The callers therefore sought "primary access" (Van Maanen and Kolb 1985) to administrative gatekeepers such as office managers, secretaries, nurses, and receptionists who control "secondary access" to physicians, the population of interest. Enlisting the cooperation of these nonphysician gatekeepers was a pivotal step in recruiting the physicians.

Table 8: Sampling Results, Cardiopulmonary Study

Source	American Medical Association Masterfile		530
	Supplemental list of Boston-area physicians		481
Disposition	Participants		192
	Eligible refusals		18
	Unreachable		5
	Ineligible:		
	Specialist, not general practice	287	
	Practice-setting stratum already filled	241	
	Not providing direct care	72	
	Retired	62	
	Racial minority	53	
	Other (e.g., moved out of area)	81	
	Total ineligible		796
Total sample frame			1011

An important issue in approaching gatekeepers was how fully they should be informed of the purpose of the contact (Becker and Meyers 1974). In this study a two-tiered approach was employed, using one of the co-investigators–a respected physician–in tandem with the research staff. The potential respondent was first mailed a personal letter from the co-investigator describing the general goals and procedures of the study and telling the addressee to expect a call. One week later, a research staff member contacted the office gatekeeper, referring to the co-investigator's letter. Specific measures were taken to distinguish this approach from pharmaceutical marketing efforts, but callers did not volunteer the information that a research subject was sought. Access to the physicians was easiest in cases where the gatekeeper had already been informed by the target physician; appointments were scheduled with a single telephone call for nine respondents who had received the letter and had asked the gatekeeper to schedule an interview or immediately forward the expected call. The importance of the advance letter was demonstrated by the fact that 13 other participants said they had not received the letter and insisted on seeing it before continuing discussions with the research staff. The greatest difficulty was experienced in practices that employed both a primary gatekeeper, who triaged all telephone calls, and an administrator, who then handled nonmedical inquiries. In these cases the research staff was obliged to inform both gatekeepers selectively about the study in order to gain access to the physician.

The distribution of telephoning effort required to set up each interview is detailed in Table 9. After at least five unsuccessful attempts to reach the physician respondent during normal work hours on different days of the week, the researcher attempted to call in the early morning (beginning at 7 a.m.) and evening (until 10 p.m.). On average, 3.4 calls were made to eligible physicians before either initial contact was made or proxy screening through

Table 9: Telephoning Effort, Cardiopulmonary Study

		Average Calls Required		
	n	*Gatekeeper*	*Physician*	*Total*
Participants	192	3.4	2.6	6.0
Hospital-based practice	64	3.8	3.0	6.8
Office-based practice	64	3.6	2.1	5.7
HMO-based practice	64	2.8	2.6	5.4
Eligible refusals	18	–	–	5.9
Ineligibles	796	–	–	4.2

gatekeepers was completed and an appointment scheduled. Additional phone calls were then required to persuade the physicians to participate and to reschedule missed appointments. On average, 6.0 calls were made to eligible respondents to schedule a single interview. The telephoning effort did not differ significantly among the three practice settings. Refusers required just as many calls as did participants, while ineligibles were identified earlier in the process and required fewer calls (Table 9).

Participation

Once access had been gained to potential physician respondents, they were screened for eligibility and asked to enroll in the study. Resistance to participation reflected the respondents' concerns over (1) legitimacy, (2) confidentiality, (3) autonomy, and (4) remuneration. Our strategy for addressing each of these concerns is described.

Legitimacy and Confidentiality. Issues of legitimacy and confidentiality have been cited as major reasons for the declining participation of physicians in research (Sudman 1985; Henderson 1978; Grady and Wallston 1988). In the cardiopulmonary study, legitimacy was conveyed by the co-investigator's letter of invitation. The letter cited funding from the National Institutes of Health and outlined the research objectives with emphasis on the practical value of a study on clinical decision making (Fowler 1984). The field title of the project, "Clinical Decision Making," reinforced its clinical relevance. The letter presented the study in a considerate manner ("We ask for your help . . ."), treating the potential respondent as an educated colleague with special expertise. The respondent was assured of complete confidentiality with the added promise that no names would appear in any report. During the interview, all participants co-signed a consent form that included a confidentiality clause.

Several respondents attributed their participation to their respect for the co-investigator who had invited them by letter. This fact raises a question of bias from preferential selection of the co-investigator's close colleagues, but that possibility is negated by the high response rate achieved in the entire sampling frame.

Autonomy. Stressing the potential respondent's importance is a more effective recruiting strategy than appealing for help or explaining a study's social utility, according to experimental and observational studies (Sudman 1985; Linsky 1965; Dillman et al. 1974). The letter of invitation therefore emphasized that the subject's personal participation was important, because he had been chosen as a representative member from a large list and his views

could not be replaced by those of another physician. Although respondents were urged, according to protocol, to schedule an interview during normal work hours in their own offices, some were unable to do so. In these cases the interviewers exercised considerable flexibility in scheduling in order to signal their interest in interviewing that particular respondent rather than anyone else. The result was a higher response rate, at a cost of minor deviation from protocol. Fifteen interviews were conducted between 6:30 a.m. and 9 a.m., and 34 were conducted after 5 p.m. Eighteen interviews were conducted at the physician's home to accommodate office policy or scheduling conflicts; these were distributed evenly among physicians from all three practice settings.

Remuneration. Experimental studies with physician subjects suggest that monetary incentives are effective and that payment, even at a level below the physician's hourly rate, is associated with a higher response rate (Sudman 1985; Gunn and Rhodes 1981; Berry and Kanouse 1987). The cardiopulmonary study, because of the length and complexity of the interview carried a relatively high incentive of $75. For the breast cancer study the amount was raised to $100. While falling short of compensation for income forgone, the monetary offer signified the importance attached to participation by the selected subjects. Three physicians who refused to participate in the cardiopulmonary study stated that the amount did not cover their time; none of the participants indicated dissatisfaction with the incentive.

SUMMARY

We employed a novel combination of experimental design and survey sampling, using simulated patients and "real" doctors, to study the doctor-patient interaction. The experimental method afforded several advantages over the observational approach taken by earlier descriptive studies in medical decision making. The factorial design enforced a balanced, controlled comparison with regard to each characteristic of the patient and physician, whereas observational studies are invariably plagued by confounding (e.g., of race and SES). The videotape technique allowed us to create "to order" a balanced sample of simulated patients, although we were obliged to use only white male physicians in the two studies conducted so far because Massachusetts lacks sufficient minorities and women in the pertinent specialties. The experimental design made extremely efficient use of the 128 physician interviews, providing more statistical power for making comparisons than would be available in a much larger observational sample. The exhaustive sampling efforts and persistent

recruitment techniques described here were crucial to the achievement of a high response rate, which minimized selection bias and enhanced the validity of these studies.

In a new study currently underway, female and minority physicians from several parts of the United States are being compared pairwise with white male physicians from the same localities. In this and other future studies, the videotape technique promises to provide insight into a variety of issues concerning different physicians' extra-medical responses to women, the elderly, ethnic minorities, and other special groups of patients.

ACKNOWLEDGMENTS

The authors thank interviewers Dennis Cohen, Rita McNally, and Jeffrey Stone; video production staff Carol Rainey, Betty Scharf, and Paul Van Ness; research staff Kate Woodman, Kathy Gear, Rich Eder, and Julie Newton; and Dr. Ralph D'Agostino, Dr. John Stoeckle, and Dr. Alan Goroll.

REFERENCES

Arndt, S. 1986. "Tips for Reaching Decision Makers." *National Underwriter* 90 (31): 2,31.

Ayanian, J. Z., B. A. Kohler, T. Abe, and A. M. Epstein. 1993. "The Relation Between Health Insurance Coverage and Clinical Outcomes Among Women with Breast Cancer." *The New England Journal of Medicine* 329 (5): 326–31.

Becker, T. M., and P. R. Meyers. 1974. "Empathy and Bravado: Interviewing Reluctant Bureaucrats." *Public Opinion Quarterly* 38 (4): 605–13.

Belanger, D., M. Moore, and I. Tannock. 1991. "How American Oncologists Treat Breast Cancer: An Assessment of the Influence of Clinical Trials." *Journal of Clinical Oncology* 9 (1): 7–16.

Berry, S. H., and D. E. Kanouse. 1987. "Physician Response to a Mailed Survey: An Experiment in Timing of Payment." *Public Opinion Quarterly* 51 (1): 102–114.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press.

Burns, R. B., K. M. Freund, M. A. Moskowitz, H. A. Feldman, L. E. Kasten, and J. B. McKinlay. Submitted. "Variability in Physician Management of a Woman with a Potential and Actual Breast Cancer."

Carter, G. M., A. Robyn, and A. M. Singer. 1983. "The Supply of Physician Researchers and Support for Research Training." Santa Monica, CA: The Rand Corporation.

Chu, J., P. Diehr, P. Feigl, G. Glaefke, C. Begg, A. Glicksman, and L. Ford. 1987. "The Effect of Age on the Care of Women with Breast Cancer in Community Hospitals." *Journal of Gerontology* 42 (2): 185–90.

Clark, J. A., D. A. Potter, and J. B. McKinlay. 1991. "Bringing Social Structure Back into Clinical Decision Making." *Social Science and Medicine* 32 (8): 853–66.

Cochran, W. G., and D. M. Cox. 1957. *Experimental Designs.* New York: John Wiley & Sons.

Dillman, D. A., E. Carpenter, J. A. Christensen, and R. M. Brooks. 1974. "Increasing Mail Questionnaire Response: A Four-State Comparison." *American Sociological Review* 39 (5): 744–56.

Early Breast Cancer Trialists' Collaborative Group. 1988. "Effects of Adjuvant Tamoxifen and of Cytotoxic Therapy on Mortality in Early Breast Cancer." *The New England Journal of Medicine* 319 (26): 1681–92.

Fowler, F. J. 1984. *Survey Research Methods.* Beverly Hills, CA: Sage Publications.

Freund, K. M., R. B. Burns, M. A. Moskowitz, H. A. Feldman, L. E. Kasten, D. A. Potter, and J. B. McKinlay. 1995. "Patient Factors Influence Physician Choices in Management of Breast Cancer." *Journal of General Internal Medicine* 10 (Supplement): 55.

Glantz, S. A., and B. K. Slinker. 1990. *Primer of Applied Regression and Analysis of Variance.* 302. New York: McGraw-Hill.

Grady, K. E., and B. S. Wallston. 1988. *Research in Health Care Settings.* Newbury Park, CA: Sage Publications.

Greenfield, S., D. M. Blanco, R. M. Elashoff, and P. A. Ganz. 1987. "Patterns of Care Related to Age of Breast Cancer Patients." *Journal of the American Medical Association* 257 (20): 2766–70.

Gunn, W. J., and I. N. Rhodes. 1981. "Physician Response Rates to a Telephone Survey: Effects of Monetary Incentive Level." *Public Opinion Quarterly* 45 (1): 109–15.

Haug, M. R., and M. G. Ory. 1987. "Issues in Elderly Patient-Provider Interactions." *Research on Aging* 9 (1): 3–44.

Henderson, S. R. 1978. "The Periodic Survey of Physicians: Design and Methodology." *Profile of Medical Practice* (1978): 177–86.

Hosmer, D. W., Jr., and S. Lemeshow. 1989. *Applied Logistic Regression.* New York: John Wiley & Sons.

Kasten, L. E., J. B. McKinlay, K. M. Freund, R. B. Burns, H. A. Feldman, and M. A. Moskowitz. Submitted. "Non-Medical Influences on Medical Decision-Making: The Role of Patient Assertiveness in Breast Cancer Care."

Keuch, R. W. 1987. "Getting to the Decisionmaker." *Life Association News* 82 (3): 115–21.

Kirk, R. E. 1982. *Experimental Design.* Belmont, CA: Brooks-Cole.

Linsky, A. S. 1965. "A Factorial Experiment in Inducing Responses to a Mail Questionnaire." *Sociology and Social Research* 49 (2): 183–89.

McKinlay, J. B., R. B. Burns, R. Durante, H. A. Feldman, K. M. Freund, B. S. Harrow, J. T. Irish, L. E. Kasten, and M. 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,* in press.

McKinlay, J. B., D. A. Potter, and H. A. Feldman. 1996. "Non-Medical Influences on Medical Decision-Making." *Social Science and Medicine* 42 (5): 769–76.

Samet, J., W. C. Hunt, C. Key, C. G. Humble, and J. S. Goodwin. 1986. "Choice of Cancer Therapy Varies with Age of Patient." *Journal of the American Medical Association* 255 (24): 3385–90.

SAS Institute Inc. 1988. *SAS/STAT User's Guide,* Release 6.03 Edition. Cary, NC: SAS Institute, Inc.

Silliman, R. A., E. Guadagnoli, A. B. Weitberg, and V. Mor. 1989. "Age as a Predictor of Diagnostic and Initial Treatment Intensity in Newly Diagnosed Breast Cancer Patients." *Journal of Gerontology* 44 (2): M46–50.

Sudman, S. 1985. "Mail Surveys of Reluctant Professionals." *Evaluation Review* 9 (3): 349–60.

Van Maanen, J., and D. Kolb. 1985. "The Professional Apprentice: Observations of Fieldwork Roles in Two Organizational Settings." *Research in the Sociology of Organizations* 4 (1): 1–33.

Yancik, R., L. G. Ries, and J. W. Yates. 1989. "Breast Cancer in Aging Women: A Population-Based Study of Contrasts in Stage, Surgery, and Survival." *Cancer* 63 (5): 976–81, Table 1.