Variations in Medical Practice Use:  
Causes and Consequences

by

Charles E. Phelps, Ph.D.  
Departments of Political Science and Economics  
and  
Department of Community and Preventive Medicine  
University of Rochester

and

Cathleen Mooney, R.N., M.S.  
Department of Community and Preventive Medicine  
University of Rochester

January 21, 1993

Correspondence:  
Charles E. Phelps  
Department of Community and Preventive Medicine  
601 Elmwood Avenue  Box 644  
Rochester, NY 14642  
(716)275-2191  
email: cphe@vm.cc.rochester.edu

KEYWORDS: Variations, medical care use, welfare loss, induced demand.

This research was supported by the Agency for Health Care Policy and Research, Grant R01 HS-06366.  
We gratefully acknowledge our University of Rochester colleagues Alvin I. Mushlin, MD, ScM, Nancy  
Perkins, MD, MPH, Laura Rice, MD, and Daniel Kido, MD in this research project, and for comments  
on previous versions, especially from Randall Ellis and Myron Stano.
I. Introduction.

The phenomenon of geographic variations in medical care use has drawn increasing attention over the past decade, and this resurgence has importantly contributed to the passage of new legislation that will direct increasingly large research budgets towards the study of the consequences of using various medical care interventions. (Chapter 3 in this volume, by Deborah Freund, details some of the history and purposes of some of this legislation and Chapter 10 by Lawrence Brown provides a policy perspective on these issues and alternative approaches to dealing with the problems caused by variability.) While some variability in resource use is both inevitable and desirable, systematic variations are common and not easily explained by random variability, differences in disease incidence, or by standard socioeconomic phenomena. The policy audience in Washington and elsewhere has taken substantial notice of the phenomenon (Roper et al. 1988, Bowen, 1988), yet the economics profession has seemed more eager to dismiss the variations literature as irrelevant, or else briefly to explain it away. In this paper, we deal with the various "dismissals" of the variations, and show why, even after accounting for these issues, variations remain a real and important phenomenon. Our paper contains new empirical results regarding one possible explanation of variations -- substitution between various medical interventions -- and clarifies their role in understanding regional variations.

We also develop a model of how variations can persist, based on the idea that physicians and patients must acquire and use costly information. We show how divergent "schools of thought" persist in the use of medical interventions. We present new evidence to suggest that such "schools" have limited boundaries that do not extend to "all" medical practice, but extend probably only within a single specialty or even a single disease. These results have implications for the degree of aggregation that one can safely use to study variations.

Finally, based on this information-cost model, we calculate the welfare losses associated with variations. This analysis contains an important addition to the variations literature -- the measure of
variation in the use of specific procedures within hospitalizations of the same diagnostic category. The welfare losses from this type of variation appears in our preliminary analysis to be nearly as large as (and add to) the welfare losses associated with variations in the rate of hospital admissions themselves.

To place our discussion in perspective, it seems valuable to distinguish between the messages of our work with that in several other chapters in this volume. Freund (Chapter 3) argues that information (in the same sense that we discuss) "may both strengthen the market and make it more competitive..." while at the same time "may also lead to greater regulation." Brown (Chapter 10) argues that this type of information provides a source of control for regulators to use. Both acknowledge the idea that information about variations in medical care use can provide the basis for regulators to set "targets" for rates of use of medical interventions, and to detect deviant regions or individual providers.

Our perspective remains somewhat broader; we view the central questions more in the area of diffusion of information. For reasons that we can only partly explore at present, it seems as if providers of medical care understand relatively poorly the proper ways to use many medical interventions, and this lack of understanding seems to pervade the production of medical care virtually independently of the regulatory environment or even the financial and legal incentives of a health care system. Perhaps the problems associated with "keeping up" in medicine are nearly insurmountable, but the issues we address may more fundamentally involve the process of initial and continuing medical education than the functioning of health care markets themselves. With these perspectives in mind, we can begin our exploration of medical practice variations.

II. The Phenomenon of Variations

To talk meaningfully about variability in the use of medical care, one needs an appropriate measuring stick. The most common measure (and, as we show below, an economically meaningful measure) has been the coefficient of variation (COV), the ratio of the standard deviation (σ) of the
distribution of rates of use divided by that distribution's mean (μ). Here, σ measures how "spread out" the pattern of use rates actually is, while μ provides the "yardstick." The COV also has the advantage that one can readily determine statistically significant differences from "no" intrinsic variability (Diehr et al, 1990, 1992). Traditionally in the relevant medical literature, "low variation" means a COV on the order of 0.1 to 0.15, and "high variation" means a COV of 0.4, 0.5 or higher. For two distributions of the same mean, a COV of 0.4 has a variance (σ^2) 16 times as large as one with a COV of 0.1, and a COV of 0.6 implies a variance 36 times larger.

The study of medical practice variations began in 1938, when Sir Alison Glover, a British physician, read a paper before the evening meeting of the Royal Society of Medicine. In that work, he showed that the rates at which school children had their tonsils removed varied greatly across regions of Britain. Using Glover's data, we have calculated the cross-county coefficient of variation (COV) in the tonsillectomy rate as .66. A few intervening studies (Glover, 1948, Lembke, 1952, Lewis, 1969, Wennberg and Gittelsohn, 1973) expanded the studies to a handful of surgical interventions, and several international comparisons of variations also appeared (Vayda, 1973, Dyck et al, 1977). The rate of appearance of variations studies increased rapidly since then, and by now, the study of variations has become a small industry, and the existence of variations seems well-documented and stable across both time and national boundaries.

We summarize this literature as follows: For some procedures where the reasons for medical intervention seem very clear (e.g., hip fracture, acute myocardial infarction), cross-regional variations are always quite low, with COVs ranging from .1 to .15 or even lower. Other interventions routinely show high variability across regions (e.g., hospitalization for back injury, back surgery, diabetes, hypertension, gastroenteritis, otitis media, etc., and several psychiatric admission categories), with COVs in the range of .4 to .6 or larger. Variations appear larger for non-surgical hospital admissions than for most surgical admissions (Phelps and Parente, 1990, Phelps and Mooney, 1992, Wennberg, Freeman and
III. "Explanations" for Variations.

The study of variations was not an hour old when the first attempt to explain away the phenomenon appeared: In discussion of Sir Alison Glover's 1938 paper on tonsillectomy variations, one physician in the audience said that:

....some of the strange facts presented by Dr. Glover...might be explained by a psychological factor — namely, maternal anxiety.... This factor alone was sufficient to explain the higher incidence...in boys than in girls and in well-to-do as compared with the poorer classes.

Since then, the attempts to dismiss variations have become more sophisticated. We summarize them here, and show why they cannot meaningfully explain the observed variations.

Economic Factors – Demand Side.

Most studies of variations adjust for the age and sex composition of the populations under consideration, but do not control for other differences in such variables as income and price. (These studies presume that the age and sex mix of a region importantly determines aggregate illness rates.) While a few studies have used direct controls for economic variables, little is known about their effect on variations in medical care use.

The ability of economic variables to explain medical care variations depends in part on the variability of those economic variables across regions, and these are not large. Intuitively, even if income or price "mattered" considerably in determining quantities demanded, these economic forces could not account for cross-regional variability in medical care use if the level of income and prices (insurance coverage) were relatively similar across regions.

In actual fact, these traditional economic variables do not differ much across regions. Per-capita
income across New York State counties has a COV of only .2 (data from NY State Yearbook). Similarly, the fraction of various areas' population covered by health insurance varies by only a little. Across counties in NY State, the proportion of hospital admissions without any insurance coverage (private or government) has a COV of only .026 (calculations available from authors). With some arbitrary assumptions about the degree of coverage offered by these plans, we can calculate the COV in net price across counties at about .08.¹ Similarly, insurance coverage varies only a little across states. With the same assumptions as before, the estimated COV in net price to consumers across states is .11 (data from HIAA, 1989, Table 1.2). The cross-state estimate is higher than the previous within-state measure because Medicaid eligibility varies across states.

How does variability in income or price affect observed variability in the use of a medical intervention? The change in the COV in a dependent variable (y) is at maximum $|\eta|$ times the change in the COV of an explanatory variable (x), where $\eta$ is the elasticity of y with respect to x (proof in appendix). This provides an upper bound on the ability of income, price, and other variables to explain COVs in medical care use rates.

Keeler et al. (1988) estimate the income elasticity of demand for medical care, particularly for hospital episodes, which we summarize as about .1. Thus, variability in income should add to the COV in use of hospital use by approximately $1 \times .2 = .02$. The price elasticity of demand for hospital care has been estimated at -.1 to -.2 (Manning et al, 1987, Keeler et al, 1988). Thus, the cross-region COV in net medical care price (COV = .08 to .11) will increase the hospital-use COV by about .016 to .022 using an estimate of $\eta = -.2$. Combined, these estimated income and price effects should add to an underlying COV in medical care use across regions by about .04. Thus, income and price variability cannot explain more than a small fraction of the total observed variability in medical care use, particularly for high-variation activities (COV = .4 to .6 or higher).

Other data show the lack of importance of income and price as major contributors to variations.
Most notably, similar patterns of variations occur both in countries with national health care systems such as Canada and Great Britain, where the marginal money costs to patients are zero, and in the United States, where non-zero prices exist. This fact alone seems to refute the idea that "standard" demand side phenomena account for the variations in any comprehensive way. Also, within the United States, the very same insurance coverage applies across diseases where variations range the entire spectrum of observed values, e.g., from .1 or smaller to .8 or larger.

Economic Factors — Supply Side.

Two major arguments exist about supply-side factors as an explanation for variable rates of use of medical interventions. The first idea depends strictly on a classic "inducement" model (patients' demands shifted for pecuniary advantage) wherein different rates of medical use depend on different rates of inducement. The second depends on an availability idea — procedures are done where the facilities exist to undertake them.

Consider first how variations might be caused by inducement. For this to hold, either the market opportunities for inducement must differ across regions, or else doctors must be willing to exploit inducement opportunities to a different degree in different regions. Neither seems plausible on its face. The same economic forces that alter normal demand for care also would alter opportunities for demand inducement, and much variation remains even after controlling for such variables. The idea that different regions have different degrees of equilibrium inducement seems greatly at odds with the standard literature on inducement, which emphasizes that demand inducement will occur at the maximum possible rate by maximizing physicians. Finally, the presence of similar variations in Great Britain and elsewhere as in the US presents a considerable embarrassment to those who would explain away variations as an inducement phenomenon, since no incentives exist to induce demand in the British or similar systems.

A second hypothesis says that the availability of beds (or other specialized resources) accounts for the variations. The evidence is mixed, and the problem of inference looms large: does demand
follow supply, or conversely? The standard "Roemers’ Law" model says that demand follows supply, while the standard economic interpretation says that supply follows demand. Given the latter possibility, even finding a positive correlation between beds and hospital use rates would not confirm causation. Some studies find positive correlations between surgery rates and the number of surgeons and/or hospital beds in a region (Wennberg and Gittelsohn, 1974, Lewis, 1969) while others show no relationship (Vayda, Mindell and Rotkow, 1982, Roos, Roos and Henteleff, 1977).

More troublesome for a straight "Roemer’s Law" interpretation, even within a single specialty such as cardiology or orthopedics, where relevant per-capita resources are held constant, large differences in regional variations exist across procedures. For example, in cardiology, the COV for acute myocardial infarction hospital admissions is about .1, but for "chest pain" or "angina pectoris" hospitalizations (relatively less severe and more poorly defined ailments) is as high as .3. In orthopedics, hospitalization for hip fractures has only small variability, but COVs rise to much higher levels for hospitalization for fracture of the ankle and fracture of the forearm, in part because more alternatives exist, e.g., ambulatory surgery or direct treatment in an emergency room (Wennberg, 1987). Since essentially the same resources combine to produce these different treatments, their presence or absence cannot explain regional differences in rates of treatment for specific diseases or using specific interventions.

**Intrinsic Differences in Illness Patterns.**

It is obviously important when comparing inter-area variations in medical care use to know the underlying patterns of illness. Much of this literature uses age and sex standardized rates of medical care use. Blumberg (1987) has challenged the idea that age adjustment accounted for much variation in health status, showing considerable variations across cities in direct measures of illness (such as bed-disability days) even after adjusting for age and sex mix. He does not, however, show how these illness differences affect hospital use, if at all.

Wennberg (1987) responds by showing that hospital use patterns are unrelated to those illness
differences (see also Wennberg and Fowler, 1977). We have estimated a regression using Blumberg's own data that shows only a very weak association between bed days and hospital days (t = 1.18) even though about one bed day out of every six is a hospital day. When we remove the hospitals days from the bed-day measure, the remaining "home-bed days" are completely uncorrelated with hospital days (t = .19, R² = .0013). Thus, one direct measure of illness (bed days) has no association with aggregate rates of hospital use. However, this study focuses on broad patterns of illness and medical resource use, not disease- or procedure-specific approaches as common in the variations literature.

A study by Roos and Roos (1982) also addresses this question. Using data on a 5% sample of the elderly in Manitoba, they assessed both the health status characteristics of individuals and rates of surgery in 56 rural regions. Their measures included self-reported health status, disability measures, and mental status, as well as standard socio-economic variables. Overall age/sex adjusted surgery rates varied by a factor of 2.7 from high to low regions; elective surgery rates varied by a factor of 3.2; cataract surgery rates varied by a factor of 4.2. Strikingly, none of the rates of surgery were more than very weakly correlated with any of the health indicators, and in most cases, in the opposite direction to "need" models of medical use.³

Neither of these studies have a health indicator that directly corresponds to a specific medical procedure. Rather, they use "general" indicators of illness. Only a small handful of studies provide such direct tests where the illness indicator focuses on a specific illness. One study of tonsillectomy in Manitoba (Roos, Roos and Henteleff, 1977) found no relationship between tonsillectomy rates and the prevalence of respiratory morbidity in a pediatric population. Using their data, we have estimated an elasticity of tonsillectomy surgery rates with respect to their measure of respiratory disease of .4, (t = .96), so while a positive relationship exists, it is very weak, and normally we would not indicate it as differing from zero.

Another study directly measuring something closely akin to "fixable illness" finds little or no
association between such illness rates and rates of several interventions. In this study, Leape et al. (1990) used reviews of medical records in 23 regions to measure the appropriateness of 3 interventions (carotid endarterectomy, upper GI endoscopy, and coronary angiography). They found substantial variations in the underlying rates of use of each of these three procedures (COV = .49 for angiography, COV = .41 for endarterectomy, COV = .21 for endoscopy). Rates of inappropriate use ranged from 8% to 75% for coronary angiography, 0 to 67% for carotid endarterectomy, and from 0 to 25% for endoscopy. For coronary angiography, inappropriate use accounted for 28% of the variations in use, but for the other two procedures, no significant correlations appeared ($R^2 = .03$). They conclude that "little of the variation in rates of use of these procedures can be explained by inappropriate use." Given their definitions of inappropriateness, this provides a good test of how illness rates affect variations; they found little if any relationship for these procedures.

**Random Noise.**

Several authors have noted that some of the data used to demonstrate variability in medical care use has potentially dubious statistical support. First, some variability would exist just due to random chance, even if each observed region produced observations from the same underlying process. The smaller the region under study, the more likely this will occur. In addition, some of the variations literature uses the ratio of highest to lowest observations; the distributions of this "extremal quotient" is unstable when small numbers of persons reside in each area under study.

One approach to assessing the importance of this phenomenon assumes that the event under study (e.g., hospitalization) occurs as in a Poisson process (McPherson et al, 1981). This allows the calculation of the "systematic" component of variation (SCV) that remains independent of chance events. In 8 surgical procedures they studied, the random component was only 1 - 4 percent of total variance in Canada, and averaging about 15 percent in England and Wales (McPherson et al, 1981, Table 3). Diehr et al. (1990) have also studied the distribution of the coefficient of variation in "typical" variations data,
and provide methods to estimate the 95% confidence intervals for coefficients of variation under the null hypothesis of equal rates of use in each region. Diehr et al. (1992) prove the relationship between a population-weighted COV measure and the chi-square statistic, thus providing precise tests for the significance of COV measures. We have applied this test systematically to our studies of variations across counties in New York State, and find that virtually all of our measures of variability, even after using regression models to adjust for age, gender, income, employment, and other differences across regions, differ significantly from zero. We are not dealing with random noise.

**Physician Differences.**

Scattered evidence suggests that individual physicians develop a practice style that predictably alters resource use, independent of patient illness characteristics. These patterns may depend upon the type of training a physician has received. Some of this evidence (Daniels and Schroeder (1977), Pineault (1977)) looks at small numbers of doctors within a single HMO. Other work (Kissick et al. (1984)) looks at doctors' decisions on a set of "papers" patients regarding doctors' decisions regarding hospitalization.

Several studies have directly looked at individual doctors' propensities to use hospital medical resources. The first of these (Wolfe and Detmer, 1984) found systematic differences across doctors on LOS in two particular operative procedures (appendectomy and herniorrhaphy). McMahon and Newbold (1986) studied length of stay within DRG groups, showing that physician practice patterns accounted for more variation than did severity of illness.

NP Roos and her colleagues have made three important contributions to this literature using Manitoba data, and "assigning" patients to doctors on the basis of frequency of contact in the previous two years. One study (Roos, 1984) identified "hysterectomy-prone" OB/GYN doctors, and compared the relative odds of a patient of such doctors having either hysterectomy or a dilatation and curettage (D&C). The relative risk of a hysterectomy for such doctors' patients was 1.6, and was 1.8 for a D&C.

A second study (Roos et al, 1986) compared the use of hospital resources for primary care
patients by doctor, standardizing for case mix. They showed that 70 percent of variation across physicians in per-patient hospital use was systematic. Further, they found negatively correlated ALOS and admission rates ($r = -.38$). Case mix adjustment removed some, but not all of the variations. Specialization of the doctor had little apparent effect on hospital use, but experience and teaching hospital affiliation reduced hospital use, as did larger numbers of primary care patients in a doctor’s practice.

Most recently, Roos (1989) calculated doctors’ propensity to hospitalize patients in one sample, and then estimated the risks of hospitalization in another independent sample of those same doctors’ patients. Using multiple logistic regression, she calculated the relative odds of hospitalization as 2 to 3 for patients whose doctors were medium and high on the overall propensity to hospitalize, compared with patients of those doctors who had a low propensity to hospitalize.

In a study still underway, Phelps, Mooney, Handy, Mushlin and Perkins (1992) have shown strongly significant differences across primary care physicians’ resource use for their patients. This study aggregates across all illnesses (of necessity, to obtain sample sizes sufficient to detect differences across doctors), and uses two separate measures to standardize the case mix and illness severity of patients across doctors’ practices. In that study, for example, ranked on per-patient resource use, the highest decile of physicians spent over three times as much as the lowest decile.

**Substitution in Production.**

Doctors can diagnose or treat some illnesses with more than one intervention. If so, observed variations in the use of any single one of those procedures would potentially overstate any welfare losses. Indeed, variations in any single procedure might have no welfare losses associated with them. For some patients, one approach may be preferred to another, but at a population level, we can think about combinations of any of these procedures to produce "cures" in a standard production function approach. Let $Q_1$ represent the number of cures attempted in Figure 1; combinations of therapies $T_1$ and $T_2$ can produce $Q_1$ attempted cures along the isoquant. Depending on the substitution possibilities, even small
differences in the relative prices of the two interventions could create significant differences in observed patterns of use. A uni-dimensional estimate of variability in the use of T₁ might show large variation, as would a separate estimate of the variability for T₂.

[Figure 1 here — Substitution in Production]

One could envision communities with different relative prices for procedures operating at points like A and B on the Q₁ isoquant in ways that were fully appropriate, with no welfare loss arising at all (assuming that Q₁ was the appropriate overall rate of treatment). Even if two communities faced the same relative prices, the welfare loss might only reflect the additional expense in community B (say) from not operating at the combination used in A. By contrast, seeing communities like A and C would imply a "pure" variations issue, namely that differing and unexplained rates of treatment represented mistakes on the extensive margin in at least one of the communities.

The correlation between use rates of interventions that potentially serve as substitutes provides a key test on this matter. If appropriate substitution is occurring, then large uni-dimensional variations could be observed with no welfare loss resulting. For this to hold, the competing interventions must show negative correlations in rates of use of. If positive correlations exist, they indicate something like points A and C, requiring different overall rates of treatment. Note that one could observe negative correlations and still have important differences in overall treatment rates. We next provide information about two alternative types of substitution: substitution of one form of treatment within the hospital for another, and substitution of inpatient for outpatient therapy.

**Substitution among Inpatient Treatments.** Within the hospital, we have identified a series of procedures that clearly have the potential to serve as substitutes. We analyzed utilization rates for these procedures in New York State using data from hospital discharge abstracts contained in the SPARCS data (see Appendix). These data provide considerable detail on diagnoses, procedures, and patient characteristics, including (importantly for our work) the county of each patient's residence. All of the
New York State data that follows assembles these records into county-of-patient origin, and calculates use-rates per capita for various interventions. In every case, results rely in residuals from reduced-form regressions that explain 30 to 90% of the raw variability.

Table 1 shows our basic results regarding substitution among procedures. This analysis provides the first results concerning substitution in production of health in the variations literature for more than a single procedure.\textsuperscript{6}

[Table 1 here — Substitution among inpatient procedures]

In every case we have investigated except one, the correlations are positive and usually significant at least at the .1 level. In all of these cases, the underlying illnesses and their treatments (cardiac disease, uterine disorders, back injuries) rank high on the list of apparent welfare losses when the procedures are ranked individually. The finding of positive correlations between the rates of interventions increases the welfare losses beyond those estimated in the uni-dimensional analysis. We summarize these results here:

**Low Back Injury.** In hospital treatment of low back injury, the correlation of non-surgical and surgical admissions was .20. Among those patients, the use of alternative diagnostic tests (myelogram and computed tomography — CT) was also positively correlated (.44).

**Coronary Artery Disease.** One of the highest variability procedures, and the one with the single highest welfare loss index, is coronary artery bypass grafts (CABG) a major surgical intervention. An alternative procedure commonly used is a "balloon angioplasty" (technically known as percutaneous transluminal angioplasty, or PCTA) where a catheter threaded into the coronary artery is then inflated, compressing the plaque clogging the artery. For single-vessel procedures, CABG and PCTA have a .50 correlation. For multiple vessel procedures, they have a .62 correlation. Surgeons also have the choice of conducting bypass surgery on a single blood vessel or multiple vessels. These procedures have a correlation of .70 in our data.

**Non-Surgical Heart Disease.** For patients with angina pectoris (chest pain due to insufficient
blood supply to the heart), alternative interventions include hospitalization and supportive care vs.
hospitalization with an arteriogram, an expensive and invasive diagnostic test. These hospitalizations have
a correlation of .08 in our data. For patients with an acute myocardial infarction ("heart attack"),
admissions with and without an arteriogram had a negative but insignificant correlation (-.18).

**Cardiac Arrhythmias.** For patients with selected cardiac arrhythmias, alternative interventions
include hospitalization with "watching" vs. surgical implantation of a cardiac pacemaker to stimulate the
heart electrically to beat in proper rhythm. These have a correlation of .16 in our data.

**Uterine Disorders other than Cancers.** For women with various uterine disorders, removal of
the uterus is commonly recommended. Two approaches exist, one with an abdominal incision and the
other with the surgery done vaginally. These two interventions have a correlation of .29 in our data.

**Uterine Fibroid Masses.** For women with fibroid masses on their uterus, alternative interventions
include a complete removal of the uterus (hysterectomy) or removal of only that portion with the fibroid
masses (myomectomy). These two surgical interventions show a correlation of .19.

**Cataracts.** For patients with vision clouded by cataracts, surgical removal of the lens of the eye
allows the return of vision. Two alternative approaches to this surgery, extracapsular and intracapsular
extraction of the lens, have a correlation of .33 in our data.

**Diagnosis of Stroke.** Patients hospitalized for a cerebrovascular accident (a "stroke") can receive
one of two diagnostic studies, an arch-arteriogram or carotid arteriogram. These procedures have a
correlation of .49 in our data.

**Intensive Care Days -- A Case of Actual Substitution.** We have actually found two cases where
substitution does occur -- the use of intensive care vs. non-intensive care beds for patients with angina
pectoris and acute myocardial infarction, where we found correlations of -.37 and -.64 respectively.

**Substitution between Inpatient and Outpatient Care**

An alternative form of substitution could occur between inpatient and outpatient treatment or
diagnosis. One study has assessed this possibility using insurance claims data in the Rochester, NY area. The insurance plans providing the data (Blue Cross and Blue Shield of the Rochester Area) cover about 75% of the under-65 population. Parente (1989) defined 19 service areas in this 6-county region, and calculated the rates of various ambulatory and inpatient surgery for residents of those regions. Table 2 shows his results. In many cases, the overall correlation coefficient is positive. Only for knee procedures does statistically significant substitution appear to take place, and then primarily in the urban sub-regions. As with substitution among alternative inpatient treatments, substitution between inpatient and outpatient interventions does not seem to explain a great part of the observed variability in hospital use. Rather, when two alternative sites for therapy might substitute one for another, we find more commonly find positive, rather than negative correlations in the rates of use across regions.

[Table 2 – Substitution between Inpatient and Outpatient Procedures]

Summary. It would appear from the available evidence that the extent of geographic variations can be explained only a little by standard economic phenomena such as price, income, and (using limited evidence) the distribution of illness across regions. Substitution in production of health does not seem to occur; rather, we systematically find positive correlations among procedures that could serve medically as substitutes. Thus, none of the standard explanations of the cause of variations provides any meaningful basis to dismiss them. We thus turn our attention to the one remaining concept — differences in beliefs about the efficacy of treatment and decisions about which patients should receive treatment.

IV. Towards a Model of Physician Learning about Treatment Efficacy

The phenomenon of variations seems widespread, and usual methods to explain them away cannot succeed, as our last section shows. This leaves unanswered the question of why variations exist and persist. In this section, we begin to build a model that explains the observed variations.

Consider an obstetrician who needs to decide on each delivery whether or not to perform a
The indications will differ across patients. These decision individually balance the benefits, risks and costs of normal delivery with those of a C-section. Given a specific distribution of patients’ characteristics, we can summarize an obstetrician’s beliefs about the efficacy of that procedure by the fraction of all patients delivered by C-section. The doctor wants to select the correct rate to perform the intervention. (We ignore here any incentives for demand inducement.)

The model we develop next corresponds to a simple “learning by doing” approach. Intuitively, it says that doctors “learn” an approach during their training (their “prior” belief) that they carry into their actual medical practices. This “prior” belief gets modified by observing the behavior of colleagues who may hold different beliefs, perhaps because they trained in different places where they learned different “truths.” In this simple model, doctors’ actual behavior will blend their training and subsequent experience. The “blend” of information will increasingly depend on “experience” as it accumulates over time. A mathematical version of this model follows:

Let $\theta$ define what the doctor believes as the “correct” rate for performing a procedure such as a Caesarian section, given a set of patients with particular signs and symptoms. We can conveniently summarize a doctor’s prior beliefs about $\theta$ using a beta distribution:

$$h(\theta) = \frac{\alpha - 1}{\Gamma(\alpha) \Gamma(\beta)} \frac{\beta - 1}{(1-\theta)^{\beta-1}}$$

The doctor’s prior experience will come (say) from residency training plus reading in journals, textbooks, etc. The size of the parameters $\alpha$ and $\beta$ summarize the extent of that experience. The parameters $\alpha + \beta$ summarize the “equivalent” number of events that a doctor has previously experienced, and $\alpha / (\alpha + \beta)$ represents the “correct” rate of applying the intervention as defined by prior knowledge.

If the doctor “learns” from ongoing experience and has a quadratic loss function regarding the divergence from his prior beliefs and the “correct” parameter $\theta$, then cumulative experience will move the doctor towards the community norm. If any individual doctor “samples” from colleagues’ behavior
regarding the right rate, then the sample of cumulative deliveries seen by an obstetrician will constitute additional experience beyond residency training. The relative costs of various sorts of information may make this a major source of information for most doctors.

For simplicity, suppose that a doctor values the expertise of each colleague similarly. The sample of deliveries (done by colleagues and observed by the doctor in his current community) $X_{i}, \ldots X_{N}$ grows with time, and $y = \Sigma y_{i}$, where $X_{i} = 1$ when the delivery is by caesarian section, $X_{i} = 0$ for normal delivery, and the doctor observes $N$ deliveries cumulatively during his practice. The expected-loss-minimizing solution to estimate the "right" rate of C-sections is given by:

$$w(y) = \frac{(\alpha + y)}{(\alpha + \beta + N)} = \frac{(\alpha + \beta)}{(\alpha + \beta + N)} \frac{\alpha}{\alpha + \beta} + \frac{N}{(\alpha + \beta + N)} \frac{y}{N}$$

i.e., a weighted average of the prior "rate" $\alpha/(\alpha + \beta)$ and the observed rate $y/N$, the relative weights being $\alpha + \beta$ for the prior and $N$ for the data. Thus, the Bayesian estimate behaves as if the doctor had previously seen $\alpha + \beta$ deliveries, $\alpha$ of which were C-sections, and blended that information with $N$ new deliveries, $y$ of which were C-sections. As $N$ accumulates, the weight on community observation dominates the weight of experience and knowledge acquired during a residency training. The weights should also depend upon the success of a doctor's own patient outcomes, compared with prior expectations. We have not yet incorporated this factor into our model.

Figure 2 shows the path of a hypothetical doctor's expected-loss-minimizing estimate of the correct rate of C-sections, assuming that the doctor's residency accumulated 5000 deliveries worth of experience, 20 percent of which were C-sections ($\alpha = 1000$, $\beta = 4000$), and that (open circles) 2000 deliveries are observed annually in a "new" city where the C-section rate is 30 percent, and (closed triangles) 10000 deliveries per year. With the higher rate of deliveries, convergence to the new community norm of 30 percent Caesarian sections occurs more rapidly.

[Figure 2 here]
These graphs show the main idea of the Bayesian learning model; the more the community norm differs from a doctor’s training, and the larger the experience he gains during practice, compared with the cumulative experience in training, the faster will the doctor converge to the community norm.

Of course, other sources of information will affect a doctor’s beliefs. Reading journals and attending professional meetings provide alternative sources of information, but these appear relatively expensive compared with the ability to talk to colleagues while in hospital scrub rooms, putting on the 9th green, etc. We anticipate that doctors would acquire information in a pattern such that the marginal costs of new "bits" of information equate across all sources. The above model demonstrates what happens when most of that information comes, in equilibrium, from colleagues in the same community.

Binomial event models like this are most appropriate when considering doctors’ decisions on the extensive margin, e.g., whether or not to undertake a treatment for a patient. A growing body of evidence shows that this extensive margin has by far the most importance in issues of medical practice variability. For example, see Roos et al, 1986, regarding the role of hospitalizations vs. length of stay in assessing variability of hospital use by physicians. See also the results of the RAND Health Insurance Study, where most of the effects of insurance came through the decision to enter treatment, rather than the intensity of treatment. However, this same general concept holds in any Bayesian learning model, no matter whether the underlying process is binomial (like C-sections) or continuous, as would be more appropriate for decisions on the extensive margin.

Consider the case of the intensity of use of a procedure, where a doctor has prior beliefs about the correct intensity distributed as \( n(\mu, \tau^2) \). The doctor accumulates experience from a process distributed \( n(\theta, \sigma^2) \) according to a sample \( X_1, X_2, \ldots, X_n \), with mean \( y = \frac{\sum X_i}{N} \), where the \( X_i \) are observations taken from surrounding doctor’s treatment of similar patients. Thus \( g(y | \theta) \) is the conditional density function of \( y \), distributed \( n(\theta, \sigma^2/N) \). Again for quadratic risk functions, the expected-loss minimizing estimate of the correct intensity is distributed normally with mean \( \omega \mu + (1-\omega)y \), where \( \omega = \frac{(\sigma^2/N)}{(\sigma^2/N + \tau^2)} \).
Thus, as the sample accumulates (with a doctor's experience in a community), the weight on the prior mean \( \mu \) (presumably formed during residency training) diminishes, and the weight on \( y \), the community norm, increases asymptotically towards 1 as \( N \) grows large.

These ideas were neatly summarized by a physician from New Haven, CT, commenting on a study (Wennberg et al, 1987) in *Lancet* that showed Boston residents were much more likely than New Haven residents to be hospitalized:

"The academic flavor in Boston, the teaching atmosphere, has a much stronger tradition of bringing people into the hospital ....[When Boston-trained physicians relocate in New Haven] they bring their bad habits with them, but peer pressure changes that."^9

Physician location choice and migration will further reinforce these phenomena. Many physicians locate into a new community by joining an established practice partnership. Selection effects by both parties will probably ensure a similar approach to medical intervention among the established doctors and the new entrant. Even those entering solo practice will have some incentive to know about any predominant local practice styles before migrating, if for no other reason because "conformance" will reduce the risk of medical malpractice suits. Thus, physician migration patterns will tend to reinforce local practice patterns.^10

**Patient Beliefs.** Equilibrium rates of treatment depend not only on doctors' but also on patients' beliefs and values. Patient's beliefs about the efficacy of an intervention may be shaped as much or more by "local" sources than will be doctors' beliefs. If a patient "doctor-shops", then any "community" style will tend to support the beliefs of the patient's own doctor. Further, if the patient samples from other similar patients in the same community to learn about their therapy, those patients' experiences will also reflect the style of the community. Thus, a community norm about medical interventions will influence not only doctor's recommendations but patients' willingness to accept those recommendations.^11

**Why Do Teaching Centers Differ in Beliefs?** The above model rests on the presumption that medical schools provide different versions of "truth" to their students. If each student received the same
information about the efficacy of a medical intervention, then each resident would hold the same priors about the efficacy of an intervention. In such a world, no variations would exist. How do such differences emerge and persist?*

The question seems odd on its face, yet if posed in economics, rather than in medicine, no economist educated in the western world would deny the importance of "schools" of thought. The "Chicago School" is perhaps most famous, with an emphasis on the market as a useful mechanism to allocate scarce resources, and with parallel emphasis in macro-economic on the "rational expectations" model. Obviously, not all universities teach the same "school" of economics. Others emphasize regulation and market failure at the micro level, and Keynesian models in macroeconomics. Schools of thought clearly exist in economics; it should not come as a surprise that they also exist in medicine.

In medicine, the opportunity for "schools" to arise comes even easier than in economics, because the theory and research underpinnings of much of medicine are specialized by organ systems (nervous, musculo-skeletal, endocrine, digestive, etc.). Thus, an orthopedic department may be "aggressive" school in using arthroscopic knee surgery, while in the same medical school, an internal medicine department may cultivate a conservative approach to treating cardiac disease. No logical reason exists for "aggressive" approaches to correlate strongly across specialty, although they may within specialty, since an overall "style" of approach to using interventions may pervade a single specialty at a particular medical center. "Schools" can arise so easily in medicine primarily because few — astonishingly few — interventions used in modern medicine have ever been subjected to carefully controlled scientific trials at a clinical level (outside of the laboratory).

The reasons for this lack of scientific support are numerous. Prominently, if an intervention (for whatever reason) has entered common practice, it becomes unethical to conduct a randomized trial on that intervention, because conducting such a trial would involve withholding the procedure for some patients. The only realistic source of funding for such trials, the National Institutes of Health, will not fund studies
that involve withholding of procedures in common use.

In addition, from an economic perspective, the normal functions that include good and exclude bad commodities in the marketplace do not exist in much of the domain of medicine: No property rights exist for a good surgical technique, nor to an excellent plan for treating diabetics. Further, no systematic liability exists for the damage wreaked by a bad surgical technique or treatment strategy, since each medical malpractice case stands on its own in the law. Thus, the incentives to produce information about the efficacy of an intervention differ considerably from those we normally find in a well functioning market for commodities. In this setting, haphazard evidence can have strong sway, creating local pockets of enthusiasm or dismay for a particular intervention. Proponents and opponents of a particular medical intervention (neither side commonly having strong clinical evidence for their beliefs) can dominate the local approach in both a medical school and in various practice communities. The Bayesian learning approach set forth above shows how, once established, such pockets of belief can and will persist.

**How Comprehensive are "Schools" of Thought?**

One common suggestion about patterns of medical care use says that "supply creates its own demand," which, if true, has important welfare implications and leads to a specific set of regulatory and policy recommendations. This viewpoint predicts that high bed supply in a region (relative to population) will uniformly lead to increased use of medical care across all diagnoses and therapies. The alternative that we suggest here says that schools of thought within a given community do not necessarily arise because of specific resource availability, and may be unrelated across various areas of medical intervention. High rates of hip surgery may well not mean high rates of knee surgery, even though both would use the same sets of resources in general. Thus, the patterns of high and low use of various medical interventions can help distinguish between a model of "Roemer's Law" vs. a model of limited information diffusion.

The patterns of medical care use formed by "schools" also affect the way analysis of medical
practice variations should take place, in particular, the desirable level of aggregation across procedures. It is easy to show that the COV of the sum of a series of partly correlated random numbers falls as the number of summed variables rises, and it falls faster, the smaller the correlations. Thus, aggregation across a large enough number of procedures can eventually produce a trivially small COV in the aggregate use rate. We can safely aggregate procedures for variations analysis only when correlations are high across procedures. Studies of variations that aggregate across procedures run a considerable risk of producing misleading results.

A study by Wennberg, Freeman and Culp (1987) suggests that the "Roemer's Law" model may have only limited applicability. Their study compared the relative use of various medical and surgical interventions in Boston and New Haven, two cities where hospitalization is dominated by faculty from prestigious medical schools. Overall, Boston uses much more hospital care on a per-capita basis than New Haven. However, for major surgical procedures, the two cities show no apparent patterns. In some cases, Boston has greater use than New Haven, and in a comparable number of other cases, the reverse holds. The greater bed supply in Boston does not alter the overall rate of major surgery, and for specific interventions, one cannot predict readily which city will exhibit the higher rate. Wennberg et al. do show that the big differences between Boston and New Haven in non-surgical admissions mostly occur for procedures where observed variations are quite high in many studies.

We have calculated the correlation of counties' use of various types of hospitalization to assess how uniform a region's patterns of use appear. If a region is "high" in all hospitalizations, the correlations will appear large, and conversely. Our model of physician learning suggests that, if anything, correlations will be higher among activities carried out by the same specialists, but will likely have little if any correlation across specialties.

[Table 3 – Correlations of Variations Across Procedures]

Our results in Table 3 provide reasonable support for the belief that the use rate across various
procedures is only weakly correlated. Medical and surgical treatment of back injury are essentially uncorrelated. Pacemaker insertions are completely uncorrelated with the two other types of heart disease admissions (acute MI and CABG), although most of the significant positive correlations occur with two types of admissions showing correlations with other procedures. These correlations may merely demonstrate inadequate adjustment (in our regression models) for the age mix of these counties.

Testing the Model. The above model offers several refutable hypotheses that we can use to test its usefulness. While we do not undertake these tests here, a research project, now underway, offers the capability of testing at least some of these ideas. This project uses data from a large IPA that assigns each patient to a primary care doctor, identifying with certainty the relationship between patients and doctors. This allows calculation of a doctor’s rates of use of various medical resources. From that basis, we can test the information model in various ways. For example:

- The deviation of a doctor’s propensity to treat from the community norm should diminish with a doctor’s time in the community.
- Doctors who trained in the community should begin their practices closer to the community norm than those who trained elsewhere.
- Doctors with more extensive specialty training should move towards the community norm at slower rates than those with less extensive training.
- Hospital-specific norms should develop just as do community norms.
- Patterns of use will correlate better within specialties (e.g., back surgery and knee surgery) than across specialties (e.g., hysterectomy and cataract removal).

We hope to be able to report the results from these types of analyses over the coming several years in our new research project.
V. Welfare Losses from Variations

If we accept the idea that variations in medical practice depend on incomplete information, then we can calculate the welfare loss associated with those variations. The basic idea says that the provision of information would alter people's behavior. Their demands when "misinformed" lead to too much or too little consumption, and commensurate welfare losses. When compared with the "right" rate of use of an intervention, communities with "not enough" use miss out on the use of some interventions where the benefits exceed — perhaps greatly — the costs. This creates one source of welfare loss. For communities with "too much" use, the costs exceed the benefits for some of the procedures performed, thus creating additional welfare losses to consumers. The "fully informed" demand curve provides the metric with which we calculate welfare losses — they allow calculation of the "right" rate.

To see how we calculate welfare losses, look at Figure 3. Suppose two communities (A and B) with otherwise comparable populations exhibit aggregate marginal value functions \( V(m) \) for a particular medical intervention according to \( V_1 \) and \( V_2 \) respectively. Suppose that this procedure has constant per unit cost \( C \), and that the differences in demand arise because of differences in beliefs about the efficacy of the intervention. The communities will exhibit (respectively) consumption rates \( m_1 \) and \( m_2 \) with optimizing behavior, given the beliefs in each region. Suppose further that the average rate of consumption \( M = (\sum X_i/N) \) is the rate that each community would use if fully informed. We relax both of these assumptions momentarily. The welfare losses are the triangles A and B, of size \( L_i = .5(X_i - M)(V(X_i) - C) \). We can estimate \( \Delta V = (V(X_i) - C) \) using information from demand studies: \( \Delta V = (X_i - M)dV/dX \). Thus, the welfare loss \( L_i = .5(X_i - M)^2dV/dX \). Adding up these welfare losses across all communities shows that \( WL = .5 \sum (X_i - M)^2dV/dX = .5N\text{Var}(X)dV/dX \) for \( N \) communities. Algebraic manipulation shows that \( WL = .5 \text{COV}(X)^2NCM/\eta \), where \( \text{COV}(X)^2 = \text{Var}(X)/M^2 \) and \( \eta \) is the elasticity of demand evaluated at \( M \). Thus, the welfare loss is proportional to the product of (a) total spending \( (N \times C \times M) \), (b) the squared coefficient of variation, and (c) the inverse demand
elasticity.\textsuperscript{18} Of course, the COV should reflect systematic, not random variability; we use regression analysis to make this adjustment. This provides a method to estimate welfare losses due to variations in medical care use, under the assumption that the variations are due to informational differences across regions. We only use unexplained residual variance in our measure of the welfare loss; the appendix in Phelps and Parente (1990) describes the precise relationships we employ here.

In these estimates, we arbitrarily assigned all residual variation to "disagreement" among physicians, and calculated the welfare loss accordingly. While this is clearly incorrect -- some variation is desirable because of variability in some unmeasured explanatory variable -- we believe that the evidence we present below confirms the economic importance of "variations" even if we discount a substantial fraction of the welfare losses that we calculate here.

**Welfare Loss with Current Average Use Too Large or Too Small.** This analysis also assumes that the average rate of use is "correct." We next prove that the welfare loss is larger if the average rate of use is biased away from the true welfare-maximizing mean. Suppose that instead of the average rate $M$, the "correct" rate is $X^* \neq M$. Then we would have overstated the welfare loss $A$ for region 1, and understated it for loss $B$ in region 2 (in Figure 3). Then:

$$WL = .5 \Sigma_n (X_i - X^*)^2 dV/dX$$

$$= .5 \Sigma_n [(X_i - M) + (M - X^*)]^2 dV/dX$$

$$= .5 \Sigma_n [(X_i - M)^2 + (M - X^*)^2] dV/dX$$

Thus, the added welfare loss due to a systematic bias is $.5N(M - X^*)^2 dV/dX$. Defining \( \%Bias = (M - X^*)/M \), then this added loss equals $.5(\%Bias)^2NCM/\eta \times (total \ spending) \times \%Bias^2 / \eta \) and this is true whether $M$ exceeds or is less than the optimal rate of use $X^*$. Figure 4 shows this phenomenon in a two-city world. Most likely, our health care system leads to too much medical care use, if anything, so $M > X^*$ because of the pervasive effects of insurance, which drives the marginal cost of care to near zero for many consumers, so we can anticipate some effects of bias on welfare loss in general.
Our empirical estimates of "welfare loss" due to variations could provide an estimate of actual welfare loss that is too high or too low, and we have no obvious way of determining which is the case. By assumption, we have modeled all deviations from "the norm" as due to disagreement about the productivity of medical interventions. We use regression analysis to eliminate systematic variability (including age, sex, income, etc.) but other unmeasured variables could cause "appropriate" variability. If so, we will overstate welfare losses due to variable use rates.

Conversely, we may understate the welfare loss from variations because we ignore within-region variability. Our information acquisition model suggests that doctors in a single hospital can develop a "style" specific to the hospital, in which case variations will exist within regions that we have not measured, and variability across individual doctors, even on the same medical staff, seems potentially important (Phelps et al., 1992, McMahon and Newbold, 1986, Roos et al., 1986).

Estimates of Welfare Losses Due to Variations in Hospital Admissions. We begin by estimating welfare losses due to variations in hospital admission rates, the aspect of variability that has almost uniformly been considered in the variations literature. We have calculated these welfare losses from using previous estimates (Phelps and Parente, 1990), and adding an estimate of the physician fees associated with the hospitalizations. Appendix 2 describes our data sources and methods in more detail; see also Phelps and Parente (1990). These results reflect only residual variance from regressions estimating reduced form models for each of 110 modified DRG groups. The R² in these regressions commonly falls in the .4 to .6 range, although in some cases, part of that explanation of variance comes through age and sex variables, since our data are not age-sex adjusted by normal epidemiologic methods. Estimates by Phelps and Parente (1990) using two years of data show that (through decomposition of variance) the behavior is quite stable through time; only a small proportion of the observed variance for any procedure (under 5 - 10 percent for almost every hospital admission type) is due to cross-year
variation in county-specific means; the remainder (usually 90 - 95 percent) by variation across counties.

We have also tested to determine if the coefficients of variation that we observe might be due to chance. Following Diehr et al. (1990), we simulated a large number of "years" of NY state data (using actual county populations) under the null hypothesis that every county had identical rates of use, and then empirically derived the 95%-ile cutoffs for various underlying rates of admission (assuming a binomial model generating hospitalizations). For various underlying probabilities in the binomial model, the cutoff rates are:

<table>
<thead>
<tr>
<th>Rate per 10,000</th>
<th>95%-ile Cutoff for COV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Test of Null Hypothesis)</td>
</tr>
<tr>
<td>5</td>
<td>.095</td>
</tr>
<tr>
<td>10</td>
<td>.068</td>
</tr>
<tr>
<td>15</td>
<td>.055</td>
</tr>
<tr>
<td>20</td>
<td>.048</td>
</tr>
</tbody>
</table>

The rate of 5/10,000 corresponds to some of our more rare (but important) events like coronary artery bypass grafts, low-back surgery, etc. Most of the other hospitalizations occur at higher rates. Thus, virtually all of the observed COVs in our data substantially exceed the 95%-ile cutoff for the null hypothesis (no differences in underlying rates of use) in our New York data.

We added physician costs to the welfare loss measure by using the estimated ratio of total costs/hospital costs reported by Mitchell et al (1984). The combined estimated welfare losses from variations in hospital admissions for 110 modified DRGs (which cover all hospital admissions) add up to $130 per capita in New York State. Extrapolating to the US population, the estimated aggregate annual welfare losses from variations in admissions reach $33 billion.

**Within-Hospital Procedure Variability.** The welfare losses discussed to this point have used the hospital admission as the unit of analysis. However, hospitalizations, even within the same DRG category, have considerable variability in their content. Our data from NY state hospitalizations (details available from authors) allows us to explore not only the variability of hospitalizations themselves, but
also variability of procedures undertaken during hospitalization. We report here our initial findings in this area, focusing on treatments of cardiac diseases, which have high welfare losses associated with them at the hospital-admission level.

A series of diagnostic tests is available to determine the cardiac condition of patients who have had a myocardial infarction (MI), commonly known as a heart attack. In general, these tests assess different aspects of cardiac performance, and hence are not true substitutes in the sense we discussed previously. In each of these (all labeled "for MI"), the relevant denominator in calculating variations is the population of persons admitted to the hospital with a diagnosis of MI. Thus, these are conditional rates of use of these tests, and they constitute additional variability beyond the rates of hospitalizations themselves. (MI hospitalizations have a COV of only .11). Table 4 displays our results for within-MI admission procedures.

[Table 4 here – Within-DRG Variations]

**ICU LOS.** Many MI patients are admitted directly to an intensive care unit initially, and then are transferred to routine hospital care when they have stabilized. Variability in length of stay for intensive care beds is low (COV = .17), reflecting a relatively uniform pattern of using this resource across regions. However, because of the expense involved, the associated welfare loss is over $6 million annually in NY state alone, almost matching the losses associated with hospitalization for pacemaker insertion or for other interventions like PCTA (balloon angioplasty).

**Cardiac Functioning.** Within the set of tests to assess cardiac function after an MI, variability of use seems large. Arteriograms (assessing the degree of openness of coronary arteries) have a COV of .32, and an annual welfare loss of $4.7 million in NY state. Stress tests occur only very infrequently, but with higher variability (COV = .40). Cardiac catheterization of the left heart occurs often, with about $7.6 million spent annually for post-MI hospitalized patients, and with relatively large variation (COV = .35). We estimate the welfare loss due to these variations at $4.7 million annually in NY state.
Radionuclide studies (COV = .30) and cardiac echo (ultrasound) tests (COV = .24) add an additional welfare loss of $2.0 million annually in NY state. Taken together, variability in these procedures, conditional on an MI admission, add $19 million annually to welfare losses, compared with an estimated loss due to variability in MI admissions of about $34 million. Thus, at least for this particular illness, within-hospital differences in medical resource use appear to create welfare losses of the same order of magnitude as those associated with admission rate differences across regions. We cannot yet determine whether this finding will recur in other types of illnesses.

How Big is "Big"?

In this section, we attempt to place in perspective the magnitude of welfare losses associated with variations in the use of various medical interventions that we have just estimated. Two comparisons may help place the magnitude of welfare loss in context. First, we compare variations-induced losses with the "traditional" welfare loss from insurance, first characterized by Arrow (1963), Pauly (1968), and formalized by Zeckhauser (1970). The second approach compares the welfare losses with the costs of creating and disseminating information about the "correct" patterns of use.

First consider the Arrow-Pauly-Zeckhauser type of welfare loss, the combination of risk bearing and standard welfare loss triangles. Using data from the RAND Health Insurance Study, Keeler et al. (1988) estimated this combined welfare loss for various insurance plans. On a per-capita basis in 1986 dollars, an uninsured person would face $1472 of loss, all due to risk bearing. A fully insured person would have $265 of loss, all due to "moral hazard" increased consumption, and would bear no financial risk. Many plans with combinations of $100 - $300 deductibles and coinsurance rates of 25% to 50% created welfare losses in the realm of $50 to $60 per person, the set of relatively efficient plans. This describes the magnitude of welfare loss for the aspect of medical care use that has attracted the attention of economists over the previous several decades. With this, we can contrast our estimate of $130 per capita in welfare loss due to hospital admissions variations, which ignores the possibility for bias in
current use rates, within-region variability, procedure-level variations within hospital admissions, and variations in medical care use outside of the hospital sector.

A second way to think about these welfare losses considers the costs of finding out the correct way to use a specific medical intervention, and compares that with the welfare losses associated with variability. The procedure with the largest estimated welfare loss is the coronary artery bypass graft (CABG), with an estimated US welfare loss of $0.95 billion annually, assuming that the current average rate of use is correct. Knowledge to reduce unwarranted variance could be worth as much as the annuity value of such losses. A twenty-five year annuity at 5%, for example, is worth about 15 times the annual cost. On those grounds, an investment to understand the correct way to use CABGs could create as much as $15 billion in improved welfare, plus gains from eliminating any bias that the studies detected.

Obviously, even a perfect understanding of the optimal ways to use medical interventions will not eliminate all variability, and we emphasize that some of the variability that we attribute to "disagreement" will ultimately prove desirable. Despite these concerns, these calculations emphasize that our estimates of the welfare losses due to lack of information (or the gains from eliminating such variations) could be greatly overstated, and it would still be a good idea to invest in knowledge about the correct ways to use these interventions, even if such research only partially reduced these variations. For procedures with losses like CABG, the case is easy to make. However, even hospital admissions like carpal tunnel release, with an estimated welfare loss of $0.139 per person per year, create annual welfare losses exceeding $34 million. The present value of learning when to use carpal tunnel release would exceed $0.5 billion, again assuming no bias in the current average use rate of the procedure, a 5% discount rate, and 15 years of benefit from the knowledge. By contrast, the costs of undertaking a careful study of a procedure like this appear to be in the realm of a few million dollars. Thus, even for a procedure with low estimated welfare loss like carpal tunnel release, it seems desirable to seek better information about ways to eliminate variability. For many other procedures, eliminating even a few percent of the observed
variation would create welfare gains exceeding the likely costs of conducting a careful study about the use of a specific procedure.

VI. Conclusions.

Variability in the use of various medical interventions appears to create important welfare losses. Many "explanations" of variations fail to account for much, if any of the observed phenomena. We demonstrate in this study how common socio-economic phenomena cannot account for more than a trivial fraction of the observed variations for many medical interventions. We also bring to bear new data to show that substitution of one procedure for another cannot account for the observed variations.

The remnants of this analysis seem to suggest that disagreement among physicians (and their patients) across regions of the country accounts for much of the observed variation, supporting the models espoused by Bloor (1976) and Wennberg (1984). If we accept this idea even partially, the welfare losses associated with variability are large. They considerably exceed those associated with the standard "moral hazard" and risk-bearing welfare losses associated with common health insurance plans, even if we ignore intra-region variability, within-DRG variability in the use of specific procedures, positive correlation of procedures that ought to be substitutes, the potential for systematic bias in the use of medical interventions, and the costs of variations outside of the hospital sector, which we have not estimated.

A model of physician learning suggests how these variations can persist in a given community. This model also implies that aggregation across medical interventions treated by different specialties will likely distort the understanding of the importance of variations. We have presented new evidence to show that patterns of variations correspond to specialty-specific beliefs about the efficacy of medical care. Thus, our work suggests that old models of "supply creating its own demand" poorly represent what is going on, at best.

Almost all previous studies of variations in medical care have used geographic regions, and have
focused on patterns of hospital admissions. We presented new evidence in this paper to show that variability within categories of admissions in the use of specific procedures is also important, creating large additional welfare losses in addition to those calculated using hospital admission rates alone. These studies emphasize the importance of analyzing variations at a relatively microscopic level.

Many other economic aspects of variations in medical care use remain unexplored. The origination of variations is in itself an important and unresolved problem (see Phelps, 1992 for some discussion of these issues). The proper approaches to modifying variable behavior also require further study. Some approaches to modifying behavior (e.g., those relying on incentives that doctors confront) may conflict greatly with approaches that seek physicians’ cooperation (e.g., assuming that the problem is one of transmission of information, rather than one of perverse incentives). Until we can resolve whether the problem centers mostly on incentives or mostly on information diffusion, the proper strategies for reducing variations cannot be determined (Phelps, 1993). What we can now say is that variations in the patterns of use of medical care seem to have high economic importance, and that research to determine the causes of variations, and to reduce inappropriate variations, should have high priority in research agendas of both governmental and private agencies.
APPENDIX

1. Proof that the coefficient of variation of \( y \) changes with respect to the coefficient of variation in \( x_i \) at a rate no more than \( \eta_i \), the elasticity of \( y \) with respect to \( x_i \):

Consider a model where \( y = \beta_0 + x_i \beta_1 + x_2 \beta_2 + \epsilon \). Then

\[
\text{Var}(y) = \text{Var}(\epsilon) + \beta_1^2 \text{Var}(x_1) + \beta_2^2 \text{Var}(x_2) + 2 \beta_1 \beta_2 \text{Cov}(x_1, x_2)
\]

\[
= \sigma^2 + \beta_1^2 \sigma_1^2 + \beta_2^2 \sigma_2^2 + 2 \beta_1 \beta_2 \sigma_{12}
\]

If we scale the entire expression by \( 1/\mu_Y^2 \) (where \( \mu_Y \) = the average value of \( y \)), then \( \sigma^2/\mu_Y^2 \) is the coefficient of variation of the residuals squared (scaled by the average value of \( y \)). Define the coefficient of variation of \( x_1 \) (\( \sigma_x/\mu_x \)) as \( C_{x_i} \eta_i \) as the elasticity of \( y \) with respect to \( x_i \), and \( \rho \) the correlation coefficient between \( x_1 \) and \( x_2 \). Then the variance expression becomes:

\[
C_Y^2 = \frac{\sigma_e^2}{\mu_Y^2} + \frac{\sigma_1^2 \eta_1^2}{\mu_1^2} + \frac{\sigma_2^2 \eta_2^2}{\mu_2^2} + \frac{2 \eta_1 \eta_2 \sigma_{12}}{\mu_1 \mu_2}
\]

\[
= C_e^2 + C_1^2 \eta_1^2 + C_2^2 \eta_2^2 + 2 \rho \eta_1 \eta_2 C_1 C_2
\]

or

\[
C_Y = (C_1^2 \eta_1^2 + C_2^2 \eta_2^2 + 2 \eta_1 \eta_2 \rho C_1 C_2 + C_e^2)^{1/2}
\]

Then

\[
dC_Y/dC_1 = \frac{\eta_1 [C_1 \eta_1 + \rho C_2 \eta_2]}{(C_1^2 \eta_1^2 + C_2^2 \eta_2^2 + 2 \eta_1 \eta_2 \rho C_1 C_2 + C_e^2)^{1/2}}
\]

\[
= \eta_1 \left[ \frac{C_1^2 \eta_1^2 + \rho^2 C_2^2 \eta_2^2 + 2 \eta_1 \eta_2 \rho C_1 C_2}{C_1^2 \eta_1^2 + C_2^2 \eta_2^2 + 2 \eta_1 \eta_2 \rho C_1 C_2 + C_e^2} \right]^{1/2}
\]

\[
= k \eta_1 \quad \text{where} \quad 0 < k \leq 1
\]

Thus, the COV of \( y \) changes with respect to the coefficient of variation of \( x_i \) at most at the rate \( \eta_i \), as asserted in the text.
NOTES

1. Define a region's the net price as $P^*$, where $P =$ actual price, $s =$ share of population with insurance, $C =$ average coinsurance for those with insurance. Then $P^* = sCP + (1-s)P = P(1-s(1-C))$. The coefficient of variation in net price is calculated assuming $C = .2$; the estimated coefficient of variation is not greatly sensitive to this assumption. Obviously, variability in $C$ across regions is not captured by this method, but this cannot add importantly to the variability in $P^*$ since most private insurance plans provide nearly full coverage for hospital care, and Medicare, which insures about one quarter of all hospital days, is uniform across all regions.

2. In the health field, this "law" is attributed to Milton Roemer, MD, who first showed the empirical relationship between hospital bed supply and per-capita hospital use. The more generic economist would refer to Say's Law.

3. For example, reports of "excellent" or "good" health came from 59 percent of respondents in areas with lowest surgical rates, and 65 percent of respondents in areas of highest surgical rates.

4. Diehr (1984) calculates that, with 1000 people in each area, the extremal quotient would have expected value of 3.2 when comparing 5 areas, and 10.9 when comparing 20 areas. Most studies of variations use many more observations than this, e.g., 40,000 to 100,000 persons, greatly reducing the problems cited by Diehr.

5. We acknowledge the assistance of Alvin Mushlin, MD, ScM, Laura Rice, MD, Dan Kido, MD, and Nancy Perkins, MD, MPH, PhD in deriving the list of procedures to consider in these tests.
6. Chassin et al, 1986 estimated the substitution of two alternative therapies for hemorrhoids within Medicare data.

7. See Hogg and Craig, 1970, Section 8.4 for proof.

8. Stano and Folland, 1988 and Stano and Folland, 1989 estimate the effects of "practice style" on use on a highly aggregated variable (all ambulatory care use). They place considerable emphasis on variations on the intensive margin of treatment (resources used per encounter), while most of the variations literature has demonstrated important variation on the extensive margin (number of encounters). They argue -- we believe in error -- that variations due to practice style should affect only the intensive margin.


10. Folland and Stano develop a model of variations on the intensive margin that requires independence of physician beliefs and location. This is clearly a risky assumption to make.

11. See Dranove (1988) for a formal model of doctors' advice-giving and patients' propensities to accept that advice.

12. This question receives more extended attention in Phelps (1992).

13. To see this, consider an aggregated variable $Y = \Sigma X_i$, with all $\mu_i = \mu$, all $\sigma_i = \sigma$, and all $\rho_i = \rho$. Thus each separate variable has a $\text{COV} = \sigma/\mu$. Then $\text{COV}(Y) = (\sigma/\mu)(N + N(N-1)\rho)^{1/2}/N =$
\((\sigma/\mu)Z\), where \(0 < Z \leq 1\) for \(0 \leq \rho \leq 1\). It is easy to show that when \(\rho = 0\), \(\text{COV}(Y) = (\sigma/\mu)/N^{1/2}\).

Also, for moderately large \(N\) (say, \(N > 10\) or \(20\)), \(\text{COV}(Y)\) is approximately \((\sigma/\mu)^{1/2}\).

14. Stano and Folland (1988) and Folland and Stano (1989) use a highly aggregated analysis of outpatient treatment. We show below that correlations across procedures and hospital admissions categories are only modestly positive, and often quite small. Thus, aggregation of hundreds of procedures is bound to mask the real variability of medical care use. To argue that variations do not exist meaningfully because of such aggregation (as Folland and Stano do) seems totally misplaced.

15. We also computed rank order correlations for the same procedures; the results show nothing different from those in Table 3.

16. The project began March 1, 1990, under funding from AHCPR.

17. Randall Ellis has pointed out to us that the marginal value rapidly falls below zero with the linear demand curve with a small elasticity (say, -.1) at the mean. However, it is easy to prove that a common alternative functional form that never produces marginal value below zero — the logarithmic transform — produces estimates of welfare loss that exceed those we estimate using the linear form, if we assume the same (constant) elasticity as we used at the average values on the linear demand curve. Intuitively, this occurs because the marginal value rapidly becomes very high for under-use with the logarithmic demand curve. Thus, our assumption of a linear demand curve may understate the true welfare losses.

18. If production costs are not constant, then an additional welfare loss arises of magnitude \(.5 \text{COV}(Q)^2\) \(\text{PQH}\), where \(H\) is the output elasticity of cost.
19. They assume a very large risk aversion measure, comparable to a relative risk aversion of about 10 or greater; other estimates suggest values in the range of 2 to 4 are appropriate. Lower risk aversion reduces the welfare losses from risk bearing proportionally. See Garber and Phelps (1992) for details.

20. The estimates from Keeler et al. refer to an under-65 population, so including an over-65 population would raise these numbers somewhat, because of the higher variance of expenditures for the elderly.
REFERENCES


Phelps CE, Parente ST, Priority setting for medical technology and medical practice assessment, Medical Care, August 1990 28(8).

Phelps CE, Mooney C, Priority setting for medical technology and medical practice assessment: correction and update, Medical Care August 1992; 31(8).


Roos NP and Roos LL, Surgical Rate Variations: Do They Reflect the Health or Socioeconomic Characteristics of the Population? *Medical Care* 1982; 20(9):945-958.

Roos NP, Predicting Hospital Utilization by the Elderly, *Medical Care* 1989; 27(10):905-919.


### TABLE 1

**CORRELATION OF SUBSTITUTABLE PROCEDURES/ADMISSION RATES**

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical back admissions and surgical back admissions, for low back problems</td>
<td>.20</td>
</tr>
<tr>
<td>Myelogram and CT, for low back problems</td>
<td>.44*</td>
</tr>
<tr>
<td>Vaginal hysterectomy and total abdominal hysterectomy, for non-malignancy</td>
<td>.29*</td>
</tr>
<tr>
<td>Total abdominal hysterectomy and myomectomy for non-malignancy</td>
<td>.19</td>
</tr>
<tr>
<td>Extracapsular and intracapsular lens extraction</td>
<td>.33**</td>
</tr>
<tr>
<td>Arch and carotid arteriogram for CVA</td>
<td>.49**</td>
</tr>
<tr>
<td>CABG and PCTA</td>
<td>.62**</td>
</tr>
<tr>
<td>SV CABG and SV PCTA</td>
<td>.50**</td>
</tr>
<tr>
<td>SV CABG and MV CABG</td>
<td>.70**</td>
</tr>
<tr>
<td>Admission for pacemaker insertions and medical admissions for selected arrhythmias</td>
<td>.16</td>
</tr>
<tr>
<td>Admission for angina with and without arteriogram</td>
<td>.08</td>
</tr>
<tr>
<td>Admission for MI with and without arteriogram</td>
<td>-.18</td>
</tr>
<tr>
<td>ICU and non-ICU LOS for MI</td>
<td>-.64**</td>
</tr>
<tr>
<td>ICU and non-ICU LOS for angina or chest pain</td>
<td>-.37*</td>
</tr>
</tbody>
</table>

**Key to abbreviations:**

- **SV**: single vessel
- **PCTA**: percutaneous transluminal angioplasty
- **CABG**: coronary artery bypass graft
- **CT**: computerized tomography
- **CVA**: cerebrovascular accident (stroke)

**MV**: multiple vessel

**LOS**: length of stay

* significant at p < .10

** significant at p < .05
Table 2

Substitution of Inpatient and Outpatient Care?

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Correlation of Use Rates</th>
<th>Inpatient vs. Outpatient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Urban</td>
</tr>
<tr>
<td>Cataracts</td>
<td>.38</td>
<td>.43</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>.44</td>
<td>.40</td>
</tr>
<tr>
<td>T &amp; A</td>
<td>.09</td>
<td>.18</td>
</tr>
<tr>
<td>Varicose Veins</td>
<td>.48</td>
<td>.81</td>
</tr>
<tr>
<td>Hemorrhoids</td>
<td>-.09</td>
<td>.13</td>
</tr>
<tr>
<td>Knee Procedures</td>
<td>-.25</td>
<td>-.72</td>
</tr>
</tbody>
</table>

**TABLE 3**

**CORRELATIONS BETWEEN PROCEDURES/ADMISSIONS RATES PERFORMED BY SPECIALISTS**

<table>
<thead>
<tr>
<th></th>
<th>CHOLE</th>
<th>HYSTER</th>
<th>CATARACT SURG BACK</th>
<th>MED</th>
<th>MI</th>
<th>CABG</th>
<th>PACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHOLE</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHOLECYSTECTOMY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HYSTER</td>
<td></td>
<td>-0.42</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HYSTERECTOMY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATARACT</td>
<td>0.17</td>
<td>0.24</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATARACT SURG.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SURGBACK</td>
<td>-0.07</td>
<td>0.07</td>
<td>0.56</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BACK SURGERY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDBACK</td>
<td>0.11</td>
<td>0.02</td>
<td>-0.07</td>
<td>0.04</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDICAL BACK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI</td>
<td>0.52</td>
<td>0.56</td>
<td>0.36</td>
<td>0.03</td>
<td>0.02</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>HEART ATTACK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABG</td>
<td>0.27</td>
<td>0.32</td>
<td>0.32</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0.21</td>
<td>1.0</td>
</tr>
<tr>
<td>BYPASS SURGERY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PACE</td>
<td>-0.03</td>
<td>0.06</td>
<td>-0.02</td>
<td>0.95</td>
<td>-0.13</td>
<td>0.04</td>
<td>-0.01</td>
</tr>
<tr>
<td>PACEMAKER INS.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Weighted by County Population

Critical values for rejecting Ho: r=0

80% confidence level: \( r > = .11 \)
95% confidence level: \( r > = .214 \)

90% confidence level: \( r > = .168 \)
99% confidence level: \( r > = .30 \)
### Table 4

**Calculation of Welfare Losses for Procedures Related to Heart Attack**

<table>
<thead>
<tr>
<th>Procedure or Admission</th>
<th>Spending Level</th>
<th>Modified COV</th>
<th>Welfare Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI (Heart Attack)</td>
<td>$496,584,013</td>
<td>11.20</td>
<td>$31,145,749</td>
</tr>
<tr>
<td>ICU LOS FOR MI</td>
<td>$42,086,475</td>
<td>17.34</td>
<td>$6,327,188</td>
</tr>
<tr>
<td>L Cardiac Cath. for MI</td>
<td>$7,682,460</td>
<td>35.25</td>
<td>$4,772,968</td>
</tr>
<tr>
<td>Arteriogram for MI</td>
<td>$9,119,090</td>
<td>32.23</td>
<td>$4,736,332</td>
</tr>
<tr>
<td>Cardiac Echo for MI</td>
<td>$4,892,736</td>
<td>28.63</td>
<td>$2,005,231</td>
</tr>
<tr>
<td>Radioisotope Studies for MI</td>
<td>$2,638,600</td>
<td>29.83</td>
<td>$1,173,951</td>
</tr>
<tr>
<td>Stress Test for MI</td>
<td>$217,562</td>
<td>46.49</td>
<td>$235,111</td>
</tr>
</tbody>
</table>

Elasticity = -1

Spending level includes both hospital and physician charges, except for ICU LOS

Key to abbreviations: MI = Myocardial Infarction, ICU = Intensive Care Unit, LOS = Length of Stay
FIGURE 2
Blending of Training Style with Community Practice