Diffusion of Information in Medical Care

Charles E. Phelps

Markets for medical care are characterized by, if not dominated by, various sorts of uncertainty (Arrow, 1963). This paper will focus on one particular uncertainty—in knowledge about the efficacy of medical interventions. Setting aside other important issues, if physicians wished to function as perfect agents for their patients, what treatments would they recommend, and how would they acquire knowledge to support those recommendations? Answering this question begins with the production of new knowledge, but extends more broadly to the diffusion and incorporation of that knowledge into patterns of treatment recommended by doctors and other health professionals.¹ The issue centers on how doctors acquire knowledge about the efficacy of medical care (its marginal productivity) and how best to move along both extensive and intensive margins to maximize patients’ health within a constrained budget. Because of the nature of medical treatment, much of the information in this category has a considerable “public good” flavor, and may be underproduced.

The paper begins with a summary of evidence that doctors behave very differently in making such recommendations, depending on the region where they work, creating large variations in the quantities of care delivered to seemingly standardized populations. This evidence on “variations” (and the failure of normal explanations of the variations) leaves almost by default the idea that incomplete diffusion of information must be largely responsible. The

¹Here the term “doctor” refers generically to any healing profession, including physicians and surgeons, dentists, chiropractors, pharmacists, nurses, psychologists, and others.

Charles E. Phelps is Professor of Political Science, Economics, and Community and Preventive Medicine, University of Rochester, Rochester, New York.
paper then discusses reasons why this problem might occur: difficulties in collecting information about the success of medical procedures; difficulties in establishing property rights to such information, even if it were to be collected; and liability considerations that hinder adoption of any information that is collected. It concludes with some suggestions for addressing these problems.

Evidence on Variations

Variations in the delivery of medical care have now been documented in many countries, using a variety of data bases and populations, for literally hundreds of medical interventions. Most of the early work on this topic focused on surgical procedures, the easiest to define and measure of all medical interventions. Almost all studies to date have used geographic regions as the unit of observation, and study rates of use of an intervention per capita. Most studies standardize the population on age and gender mix and then assume that the underlying rates of illness in the populations either are the same or have little meaningful variability.

The most common measure of variability in the literature is the coefficient of variation (COV), defined the ratio of the standard deviation to the mean of the observed distribution of per-capita rates of use of an intervention, adjusted for age and gender. Because it uses standard deviation, the COV “shrinks” variability that one would normally think of when comparing means and variances of distributions. For two procedures with the same mean, a COV of 0.4 has a variance 16 times higher than one with a COV of 0.1.

Other studies report the extremal quotient (the ratio of the highest to the lowest observed rates). As a rule of thumb, it will often hold true that the ratio of highest to the lowest rates of use will approximately correspond to ten times the coefficient of variation. If the COV for appendectomy is .30, for example, then the ratio of high/low will be roughly 3. This holds true for most procedures and studies in this literature, and may provide a more intuitive way of comprehending the meaning of a coefficient of variation number.²

The study of variations in medical practice began in 1938, when a British physician named Sir Allison Glover read a paper before an evening meeting of

²The COV has the advantage that it can readily be tested against the hypothesis that the underlying rates of use of the intervention are the same in each observed region (Diehr et al., 1992), since a scalar multiple of the COV² (when calculated with each of k units of observation weighted by its population) has a χ² distribution. Specifically, where k = the number of small areas, N = the total population across all areas, p = the rate of hospitalization, then COV²(k – 1)Np/[1 – p] is χ²(k). For small p (e.g., 1/3000), and modestly large k (e.g., 50 to 100), this is closely approximated by COV²Np. Note that Np equals the total number of cases across the entire population in a given period. Suppose k = 30, and N = 5 × 10⁵. Now consider a procedure with an underlying average rate of p = 10⁻³ (1 per 1,000). Then Np = 5000. If the COV = .1, the then approximate test statistic is 50, which is significant at a .012 level. Procedures with a COV = .4 can occur at a rate of 16 times less often and have the same level of significance.
Table 1

Coefficients of Variation of Surgical Removal Procedures in Various Studies

<table>
<thead>
<tr>
<th></th>
<th>Prostate</th>
<th>Tonsil</th>
<th>Appendix</th>
<th>Hernia</th>
<th>Hemorrhoids</th>
<th>Gall Bladder</th>
<th>Uterus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kansas</td>
<td>—</td>
<td>0.29</td>
<td>0.52</td>
<td>0.22</td>
<td>0.40</td>
<td>0.32</td>
<td>—</td>
</tr>
<tr>
<td>Northeast U.S.A.</td>
<td>0.30</td>
<td>0.36</td>
<td>0.26</td>
<td>0.11</td>
<td>0.30</td>
<td>0.18</td>
<td>0.22</td>
</tr>
<tr>
<td>Norway</td>
<td>0.33</td>
<td>0.48</td>
<td>0.16</td>
<td>0.20</td>
<td>0.47</td>
<td>0.18</td>
<td>0.31</td>
</tr>
<tr>
<td>West Midlands</td>
<td>0.24</td>
<td>0.31</td>
<td>0.16</td>
<td>0.20</td>
<td>0.35</td>
<td>0.16</td>
<td>0.20</td>
</tr>
<tr>
<td>Maine</td>
<td>0.26</td>
<td>0.43</td>
<td>0.18</td>
<td>0.14</td>
<td>0.55</td>
<td>0.23</td>
<td>0.25</td>
</tr>
<tr>
<td>England and Wales</td>
<td>0.22</td>
<td>0.19</td>
<td>0.13</td>
<td>0.16</td>
<td>0.24</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>Canada</td>
<td>0.33</td>
<td>0.23</td>
<td>0.15</td>
<td>0.14</td>
<td>0.35</td>
<td>0.14</td>
<td>0.18</td>
</tr>
<tr>
<td>New York Counties</td>
<td>0.18</td>
<td>0.42</td>
<td>0.21</td>
<td>0.16</td>
<td>0.16</td>
<td>0.14</td>
<td>0.28</td>
</tr>
</tbody>
</table>


The rates at which school children had their tonsils removed varied greatly across regions of Britain. Using his data, one can calculate the cross-county coefficient of variation (COV) in the tonsillectomy rate as .66. In another classic study of tonsillectomy, the chances of a child having their tonsils removed varied across towns in Vermont from 7 to 70 percent (Wennberg and Gittelsohn, 1973). A few intervening studies (Glover, 1948; Lembke, 1959; Lewis, 1969; Wennberg and Gittelsohn, 1973) expanded the studies to a handful of surgical interventions, and several international comparisons of variations also appeared (Pearson et al., 1968; Vayda, 1973; McPherson et al., 1981; McPherson et al., 1982).

These studies all find large variability in cross-regional rates of surgery, with strong commonalities appearing in the patterns even across countries with widely disparate health care financing systems. The rate of appearance of variations studies increased rapidly since then; 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.

Table 1 shows the coefficients of variation for a handful of common surgical interventions that have been studied repeatedly through time and across space. Commonalities appear in these studies that begin to suggest an underlying source for the variations. When the disease is very easy to diagnose, the consequences of not intervening are well understood, and few alternative interventions exist to treat the disease, then observed variability is quite low, on the order of 0.1 to 0.2. Hernia repair and removal of an inflamed appendix (appendectomy) provide two good examples. Alternatively, when the “indications” for surgery are less clear, or when alternative treatments exist (such as surgery or bed rest plus therapy for low back injuries), variations increase.
The particular procedures shown in Table 1 are not necessarily those with large coefficients of variation; rather, they are the ones reported most commonly in studies. Larger coefficients of variation have appeared in several studies for some procedures. For example, in Medicare patients, Chassin et al. (1986) found high variation in such procedures as injection of hemorrhoids (COV = .79), hip reconstruction (COV = .69), removal of skin lesions (.67), total knee replacement (.47), and others. In a study of hospitalizations in New York state (Phelps and Parente, 1990), the largest coefficients of variation for hospitalization occurred for within-hospital dental extractions (COV = .73) and false labor (COV = .75). Tonsillectomy, surgery for low back injuries, and coronary artery bypass grafts (CABG, pronounced “cabbage”) are “classic” examples of high variations procedures, with a COV often exceeding 0.4.

Several studies have also provided coefficients of variation for non-surgical procedures, and these data show that the uncertainty about hospitalization is at least as great in this area as for surgery. The first of such studies found very high variations, with COV often exceeding 0.4, in hospital admissions for relatively simple diseases such as urinary tract infections, chest pain, bronchitis, middle ear infections and upper respiratory infections (both adults and children), and pediatric pneumonia (Wennberg, McPherson and Caper, 1984). In a study of Medicare patient hospitalizations, Chassin et al. (1986) found moderate to large coefficients of variation for a number of non-surgical conditions, including diagnostic activities such as skin biopsy (COV = .58) and coronary angiography, to detect clogging of arteries into the heart (COV = .32). In New York, Phelps and Parente (1990) found large variations for a substantial number of pediatric hospitalizations, even after controlling for the age mix of the populations, including pneumonia (COV = .54), middle-ear infections and upper-respiratory infections (COV = .60), bronchitis and asthma (COV = .36), and gastroenteritis (COV = .45). Large variations also occurred for adult admissions in categories such as concussion (COV = .43), chronic obstructive lung disease (COV = .4), medical back problems (.31), adult gastroenteritis (.25) and similar diseases. Psychiatric hospital admissions were also quite variable; for example, depression (COV = .45); acute adjustment reaction (COV = .48); and psychosis (COV = .29).

In yet another recent study, Wennberg, Freeman, and Culp (1987) reported on a comparison of the use of various medical procedures in two cities in the United States (Boston and New Haven), where in both cities a large fraction of hospitalizations occur in hospitals affiliated with medical schools (87 percent in Boston, 97 percent in New Haven). One important feature of this study is that it shows substantial variations in practice patterns even within the part of the medical community—academic medicine—that should have the best information about the efficacy of various medical interventions. The cities are quite similar in terms of age profiles, the level and distribution of income, the extent of insurance coverage, and the proportion of persons who are non-white. Thus, neither income nor out-of-pocket price can explain much of the difference in utilization between these cities, especially when one notes the
quite small income and price elasticities for hospital care (Manning et al., 1987; Manning and Marquis, 1989). By contrast, Boston has 55 percent more hospital beds per capita (the extensive margin) and each hospital bed had 22 percent more hospital employees (the intensive margin) who were paid on average 5 percent more, than in New Haven. On average, residents of the Boston area spent 87 percent more on hospital care than those in New Haven.

The age-adjusted patterns of medical care use by citizens of the Boston area uniformly are higher than those of New Haven, with most of the difference occurring in minor surgery cases and in medical (non-surgical) cases where variations in admissions rates are high across the country. Small differences exist for major surgery and low-variation medical admissions, both in admission rates and lengths of stay.

Wennberg (1990) has documented similarly large variations in the admission rates for numerous surgical procedures in the market areas of 16 major university hospitals and large community hospitals around the country—again, in medical centers with the greatest presumed medical knowledge of any part of the health care community. The variations in admission rates correspond closely to those found in other settings. Even medicine's elite systematically disagree about the proper use of many procedures. Table 2 shows his findings for 30 surgical procedures. Table 2 also displays the rule of thumb that the coefficient of variation, multiplied by ten, is roughly equal to the high/low ratio.

The underlying premise of much of the literature on medical practice variations holds that cross-regional differences in the underlying patterns of disease do not differ nearly as much as the differences in treatment. In most cases, one would expect age and gender differences in populations to play an important role in the underlying patterns of disease, but almost all relevant studies of practice variations hold age and gender mix constant. In many other cases, there is no obvious reason why there should be important cross-regional differences in the patterns of disease, and in many cases, it has remained an untested assumption that the underlying frequency of disease is nearly the same in every region after correcting for age and gender differences. In some cases, the underlying disease process has also been measured directly or indirectly, and has generally been found to have little meaningful variability across region, or else it has been found that the variability in treatment was essentially unrelated to the patterns of morbidity measured. This opens the obvious and central question: If differences in populations' illness patterns do not explain the differences in medical care use, what does?

Seeking Explanations for Variations

A number of obvious reasons arise to explain why we should see variations in medical care such as have been observed, analyzed in further detail in Phelps and Mooney (forthcoming) and Handy, Phelps and Mooney (1992). These
Table 2  
Coefficients of Variation in Medical School Cities

<table>
<thead>
<tr>
<th>Surgical procedure</th>
<th>Number of cases</th>
<th>Coefficient variation</th>
<th>Ratio High / Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colectomy</td>
<td>3,190</td>
<td>.116</td>
<td>1.47</td>
</tr>
<tr>
<td>Resection of small intestine</td>
<td>1,017</td>
<td>.142</td>
<td>1.75</td>
</tr>
<tr>
<td>Inguinal hernia repair</td>
<td>9,795</td>
<td>.152</td>
<td>2.01</td>
</tr>
<tr>
<td>Pneumonectomy</td>
<td>505</td>
<td>.213</td>
<td>2.72</td>
</tr>
<tr>
<td>Extended simple radical mastectomy</td>
<td>2,012</td>
<td>.214</td>
<td>2.21</td>
</tr>
<tr>
<td>Cholecystectomy</td>
<td>8,558</td>
<td>.231</td>
<td>2.22</td>
</tr>
<tr>
<td>Open heart surgery</td>
<td>1,439</td>
<td>.282</td>
<td>2.29</td>
</tr>
<tr>
<td>Simple mastectomy</td>
<td>359</td>
<td>.266</td>
<td>2.71</td>
</tr>
<tr>
<td>Proctectomy</td>
<td>927</td>
<td>.272</td>
<td>3.01</td>
</tr>
<tr>
<td>Repair of retina</td>
<td>1,134</td>
<td>.274</td>
<td>3.12</td>
</tr>
<tr>
<td>Hysterectomy</td>
<td>10,055</td>
<td>.275</td>
<td>2.60</td>
</tr>
<tr>
<td>Pacemaker insertion</td>
<td>3,430</td>
<td>.281</td>
<td>2.63</td>
</tr>
<tr>
<td>Appendectomy</td>
<td>5,381</td>
<td>.305</td>
<td>2.86</td>
</tr>
<tr>
<td>Prostatectomy</td>
<td>6,379</td>
<td>.327</td>
<td>3.12</td>
</tr>
<tr>
<td>Thyroidectomy</td>
<td>949</td>
<td>.342</td>
<td>3.35</td>
</tr>
<tr>
<td>Total hip replacement</td>
<td>1,717</td>
<td>.353</td>
<td>2.99</td>
</tr>
<tr>
<td>Peripheral artery bypass</td>
<td>1,455</td>
<td>.359</td>
<td>4.36</td>
</tr>
<tr>
<td>Embolectomy, lower limb artery</td>
<td>529</td>
<td>.364</td>
<td>4.10</td>
</tr>
<tr>
<td>Diaphragmatic hernia</td>
<td>2,178</td>
<td>.369</td>
<td>3.45</td>
</tr>
<tr>
<td>Coronary bypass surgery</td>
<td>3,744</td>
<td>.383</td>
<td>3.62</td>
</tr>
<tr>
<td>Aorto-iliac-femoral bypass</td>
<td>551</td>
<td>.384</td>
<td>4.07</td>
</tr>
<tr>
<td>Graph replacement of aortic aneurysm</td>
<td>491</td>
<td>.402</td>
<td>6.26</td>
</tr>
<tr>
<td>Excision of intravertebral disc</td>
<td>4,240</td>
<td>.433</td>
<td>5.09</td>
</tr>
<tr>
<td>Cardiac catherization</td>
<td>9,952</td>
<td>.443</td>
<td>4.48</td>
</tr>
<tr>
<td>Mastoidectomy</td>
<td>569</td>
<td>.461</td>
<td>4.03</td>
</tr>
<tr>
<td>Laparotomy</td>
<td>4,126</td>
<td>.471</td>
<td>5.60</td>
</tr>
<tr>
<td>Stapes mobilization</td>
<td>606</td>
<td>.483</td>
<td>4.28</td>
</tr>
<tr>
<td>Spinal fusion with or without disc excision</td>
<td>1,234</td>
<td>.520</td>
<td>5.20</td>
</tr>
<tr>
<td>Total knee replacement</td>
<td>998</td>
<td>.525</td>
<td>7.42</td>
</tr>
<tr>
<td>Carotid endarterectomy</td>
<td>1,471</td>
<td>.825</td>
<td>19.39</td>
</tr>
</tbody>
</table>

Source: Wennberg (1990)

fall into three main categories: random noise; individual constraints and preferences, including standard economic effects of price and income; and substitution in production of health.

Taking first the issue of whether the variations are “noise” or real differences, one can say with confidence that virtually all of the studies of variations have a degree of variability that rules out “noise” as a sufficient explanation, although some of the observed variation is certainly noise. (Here, “noise” means the stochastic component of illness rates for single diseases: that is, cross-regional differences in observed utilization that would emerge even if the underlying true stochastic process generating illnesses were the same in all geographic regions.) The most prominent work in this area has used simulation
studies to show what sorts of variability would arise under the null hypothesis of “no variability” (Diehr et al., 1992). That work has established guidelines by which previous studies can be compared, suggesting that many, if not all of the observed studies have found “real” systematic variability.

The second issue concerns the role of individual preferences and economic constraints, including income and substitution effects. Two approaches to this problem suggest that the observed variations are not mainly due to individual preferences. First, “typical” patterns of variability have been well-documented both in the British National Health Service (a socialized health care system) and in Canada, where universal, full coverage insurance exists. In general, this rules out both price and income effects as having a meaningful role, since price is eliminated from consideration, and full insurance drives the income elasticity to zero as well (Phelps, 1976). Variations also exist in the Canadian, Swedish, Norwegian, and other health care systems where all patients have virtually complete insurance, so the phenomenon is worldwide.

Separately, it is possible to determine an upper bound for the possible effect of exogenous factors on variations between regions. Phelps and Mooney (forthcoming) show that the effect of any exogenous variable on observed cross-regional variability is limited by the variability in the underlying explanatory variable across regions, and by the relevant demand elasticities. In particular, the coefficient of variation of a dependent variable $y$ (for example, rates of low back surgery) across regions will change with respect to the COV of any explanatory variable $x$ (like income or insurance coverage) at most by the rate of the elasticity of $y$ with respect to $x$. Using population-weighted data from counties in New York state, one can obtain an estimate of the variability of income and price across regions of .2 for income and 0.1 for net price of hospital care (Phelps and Mooney, forthcoming). The respective demand elasticities were estimated using data from the RAND Health Insurance Study at about 0.2 and $-0.2$ (Manning and Marquis, 1989, for income elasticity, Manning et al., 1987, for price elasticity). Thus, the total contribution to the COV for hospitalizations due to income and price has an upper bound of about .06. Differences in the age mix can also create different patterns of use, but almost all reported studies control for age mix of populations studied.

The idea that personal preferences should create differences in patterns of medical care use has obvious validity. However, for preference differences to explain cross-regional variability in medical care use, one must demonstrate that the average preferences for health (vs. other goods) differ across regions. While individuals certainly have different preferences, the variations literature measures behavior at a highly aggregated regional level, and we have no reason to expect that preferences will differ on average across regions. Indeed, one study that directly measured cross-regional differences in attitudes that should affect medical care use (for example, belief in the efficacy of medical care) found no meaningful differences across towns in Vermont, even though they had found order-of-magnitude differences in rates of tonsillectomy across those same towns (Wennberg and Fowler, 1977).
The third issue involves substitution in production. If two alternative treatments are available for the same disease, one might find widespread differences in the rates of use of both interventions, but little overall difference in the aggregate rate of treatment of the disease. Small differences in preferences of patients or efficiency of providers could lead to substitution, and hence explain the observed variations. For example, bypass surgery (replacing clogged coronary arteries) and a procedure known as “balloon angioplasty” (inserting a catheter into the artery, inflating it like a balloon, and compressing the cholesterol plaques that had clogged the artery) offer two alternative treatments for patients with insufficient blood flow to the heart. If each was a substitute for the other, the data should show communities with high rates of one procedure having low rates of the other. However, this substitution does not seem to occur often. Phelps and Mooney (forthcoming) found that, rather than observing substitution, the use of interventions that could serve as medical substitutes were instead positively correlated.

In short, large variability does exist in average rates of use across regions, even when one holds constant age (the most important predictor of medical care use), income, insurance coverage, and other factors that would normally enter an economic model of demand. Thus, the most widely offered explanation for the variability in medical care use across regions is “medical uncertainty” (Wennberg, 1984). Using that interpretation, Phelps and Parente (1990) estimate the aggregate annual welfare loss from practice variations in excess of $7 billion, counting only cross-regional (not within-regional) variation, and only variability in hospitalization rates. Additional variation in other resource use, and within region, will add to the aggregate welfare losses from unexplained variability. Thus, even if a significant fraction of the observed variability is somehow proven “desirable,” the remaining fraction that one could meaningfully attribute to disagreement about the parameters of the production function still creates a substantial welfare loss. Indeed, the welfare loss appears to be of the same general magnitude as that commonly attributed to “moral hazard” arising from altered incentives to use medical care in our health insurance system (Phelps and Mooney, forthcoming).

The logical steps between individual medical uncertainty and widespread regional variability in medical care use are not obvious. After all, if doctors

---

3The basic method of Phelps and Parente is as follows: Consider the demand curve for any single procedure $X$, and suppose that any external intervention created a rate of use of $X$ that differed (either high or low) from the optimal consumption (the point where the demand curve intersected the marginal cost curve). A rate of use that was too high would create a welfare loss where more resources were used than value created. A rate of use that was too low would create welfare loss from foregone consumer surplus. The hypothesis in their work is that incomplete knowledge creates just such variability around the true “correct” rate, albeit unknown. The welfare loss from variability alone can then be estimated from standard welfare triangles as $WL = .5PQ(\text{COV}^2)/\eta$, where $PQ$ represents aggregate spending in procedure $X$ and $\text{COV}$ is the coefficient of variation measured around the observed mean rate of use. Additional welfare losses arise if the average rate of use is biased.
everywhere were uncertain to the same degree, they should on average behave in the same way in a region, which obviously does not occur. Something creates regional patterns of use.

One explanation centers on the problems of inference confronting doctors. In this model, when "schools of thought" get established in a specific locality, the costs of gathering relevant evidence to alter those beliefs will often be large, if not prohibitive. Thus, local "schools" can emerge and persist that hold different beliefs about the efficacy of an intervention, and these schools can in turn lead to different recommendations from doctors to their patients, and hence to different patterns of practice in the aggregate. Since medical care is generally a local service industry (with a few exceptions involving regional or national referral patterns for highly specialized surgery), one does not find usual market forces driving out inefficient modes of production (in this case, incorrect treatment strategies) with the same vigor as one finds (say) in national markets for manufactured goods.

A simple example of these "schools" arises when one looks at the average length of stay in hospitals for the various states. There are large differences in average length of stay across states and through time, but the geographic patterns suggest a wave-like transmission of "knowledge" that is quite dramatic. Figure 1 shows a "weather map" of the U.S. with boundaries showing patterns of similar average length of stay within each state. Even at this highly aggregated measure (all hospitalizations) and looking only at length of stay, it is clear that important differences exist across states, and that some patterns of commonality also appear. How this happens may depend in part on the problem doctors have in determining whether one strategy is really better than another. The next section considers this problem in more detail.

Figure 1

Days of Average Hospital Stay, By State

U.S. average = 7.1
The Inference Problem

Consider the problem of inference confronting an individual doctor, a group of doctors, or even an entire community of doctors. One of their colleagues hears of a new intervention to treat a disease, rather than the customary treatment used repeatedly through the years. They wish to know if the new treatment has better outcomes, on average, than the existing treatment, perhaps for all patients, or perhaps only for some identifiable subset. Most doctors will not be able to answer this question for most diseases based on their own patient care—the sample sizes available to any individual are too restricted to provide meaningful answers.\(^4\)

However, their attempts to answer such questions provides one basis for understanding how variations arise and persist. With relatively small samples available for “local” tests of a new intervention, natural variability in outcomes of small-sample tests can lead some communities to adopt a new intervention, while others decide not to, on the basis of small-sample inferences.\(^5\)

To see how this can happen, consider testing of a new intervention vs. an old one. Suppose doctors know that the probability of success of the old intervention is 0.6, and that they decide to adopt the new intervention in their own community if it is more successful by some statistically significant amount. They gather a sample of (say) 100 new patients, to compare against previous years’ patients. If the true rate of cure for the new treatment is 0.7 (vs. 0.6 for the old therapy), and they use a significance level of 10 percent, then they will only switch treatments 37 percent of the time. If the cure rate for the new treatment is .75, they will switch only 68 percent of the time (Fleiss, 1984). Thus, even if the new treatment is 25 percent more effective than the old one, only two-thirds of the communities will adopt it under such a testing regime. The symmetric case also holds; if the new treatment is worse than the old one, some of the communities will still adopt it because of sampling error.

Another interesting conclusion from this line of thought is that if doctors are predisposed to accept the new treatment, the minimal evidence available to them will commonly lead them not to reject the hypothesis that the new treatment is better than the old one. A number of economic and psychological

\(^4\)There are other sources of information as well, most obviously medical journals, meetings, etc. Here an important distinction arises; medical journals and the like are a very efficient way of disseminating information from larger studies. The problem of individual doctors acquiring that information is considerable. In many illnesses and treatments, literally hundreds of journals exist that could report relevant results. Yet most physicians must “keep up” in the treatment of hundreds, if not thousands of diseases. Economists, by contrast, usually have the luxury of limiting the lines of inquiry that we take—we specialize. Few economists actively conduct research in more than a few narrowly defined areas. We can limit our journal reading to a few selected journals and have high confidence that most of the relevant literature will reach them. Physicians have no such luxury, however, so the economics of using the information, even when available in published form, probably dominates the economics of producing it.

\(^5\)Alan Garber initially suggested this line of reasoning to me.
forces might lead to this choice of the null hypothesis. The comparative profitability of old and new treatments can matter. Also, doctors may gain some reputational value from being “out in front” on new treatments. Finally, doctors, like economists, may have personal preferences that lead them to form hypotheses in such fashion when they are otherwise indifferent, a trait captured in the idea of the “technological imperative” to do new things when available. If the null hypothesis is that the new treatment is better than the old, then small samples have very little power to reject that hypothesis unless treatment differences are very large.

Once having adopted a therapy, a community may well persist in using it. Even if doctors enter a community with different ideas, they may well converge on a single local approach as information diffuses between them (Phelps and Mooney, forthcoming). Legal incentives reinforce this result, for reasons explained in the next section.

How difficult is collecting large samples to learn about treatment efficacy? For many diseases, the process is amazingly difficult, because individual diseases are intrinsically rare. A “common” disease occurs in only several patients per thousand population. A city of 100,000 persons will have only several hundred patients with a specific disease in a year, and these could easily be dispersed across numerous physicians, even with a tightly concentrated referral system. More likely, any disease in question will be so rare that a city must have millions of inhabitants before the illness becomes sufficiently common to allow accurate estimation of treatments based on that single city. Thus, testing a treatment’s efficacy can take years if the data collection is limited to single cities.

Two common responses to this problem occur. The first is that local testing takes place with very small samples, usually in the hands of one or two doctors who learn of a new therapy. If it “works” for them—based on whatever standards they apply—it will likely become the standard of practice in a community; if not, the old therapy will be retained. “Testing” with small handfuls of patients will quite likely be the norm in many settings. Indeed, any systematic reading of the medical literature reveals numerous publications where a small series of “cases” are reported and conclusions drawn about the treatment’s efficacy, often without the benefit of a control group, let alone with sufficient sample size to draw meaningful conclusions.6 “Meta-analyses,” summarizing bodies of literature regarding a given intervention, must often discard large fractions of the published articles because they contain no control group, and hence offer no scientific basis by which to judge a therapy’s efficacy.

The other alternative to learn about a treatment’s efficacy is a formal randomized trial, organized across many sites and years. These studies, as funded by the National Institutes of Health, commonly involve 10 to 25 sites

---

6 The issue of small sample size in medical studies calls to mind the whimsical “report” from a study of a treatment’s efficacy: “In our study, one third of the treated animals were cured. One third died. The other one got away.”
and collect data for several years. Only under such settings can relatively precise testing of “effects” take place. In many randomized trial reports in the medical literature, the list of collaborating physicians takes more space than the “results” section of the paper, attesting to the dispersed nature of diseases and the difficulty in achieving adequate statistical power.

The inference problem with small sample size is compounded by many possible sources of statistical variation. There can be patient-specific effects (almost unmeasurable in any meaningful sense), doctor-specific effects, hospital-specific effects, and even time (experience) effects (Luft, Bunker and Enthoven, 1979; Luft, 1980). These all hamper one’s ability to measure the true marginal productivity of a medical intervention, which may in turn differ for patients of specific characteristics (age, race, gender).

The inference problem confronting doctors who wish to determine the efficacy of any specific intervention is hellishly complex, and the decision to undertake such studies will depend in large part on the incentives offered for the pursuit of such knowledge.

Is Medical Care Different?

These issues suggest that something happens in markets for medical care differently from that in other markets. In particular, one can ask how the problem of choosing whom to treat with a particular medical intervention differs from other production choices. As discussed previously, one might suspect some differences arising from the sheer number of treatment strategies that a physician must master. Some important work on the diffusion of technological innovations sheds light on this problem. Griliches (1957, 1960) has studied the adoption by U.S. farmers of hybrid corn. He shows a pattern of diffusion that plots percent of farmers adopting the new technology vs. time. In every state studied, the diffusion curve follows a classic “S” shape, with diffusion (from no use to nearly 100 percent use) taking place over a span of 10 to 20 years. Some states began the diffusion process later than others, so that (for example) Texas and Alabama began the diffusion in about 1945 only after Iowa farmers had completely accepted the new corn seed hybrid. Griliches reports that the same logistic diffusion process represents well the pattern of adoption for many types of farm equipment (across states), and across smaller sub-units (the county and crop-reporting district). Thus, if one looked cross-sectionally at any of these markets during the diffusion process, one could easily observe the same “variations” in use of a technology (like hybrid seed corn) that one observes in health care markets.

An important difference appears between these studies of technological diffusion and those in health care: In the diffusion of most technological innovations, the superior strategy eventually overtakes the market. In Iowa farm markets, the diffusion took about a decade, and in other markets, about
two decades. By contrast, in medical care markets, there seems to be no convergence through time in the use of specific interventions. Variable patterns of use of tonsillectomy were first measured in 1938, and widespread "official" pronouncements about the undesirability of widespread use of tonsillectomy have existed for decades within pediatrics specialty societies. Yet, disagreement about the proper use of tonsillectomy is nearly as high now as it was 50 years ago: the COV revealed in Glover's 1938 study of tonsillectomy use closely matches that found a decade later in England (Glover, 1948), in Kansas in 1963, and through numerous other studies to those even using late 1980s data in New York (Phelps and Parente, 1990). There seems to be no convergence in medical markets. This feature alone distinguishes the market for medical care from the diffusion of technological change in other markets. Following sections discuss reasons why we might expect to find such differences.

Legal Incentives and Property Rights in Medical Information

The role of legal incentives appears first in the way variability in medical practice is sustained, once established, and then in the production of knowledge itself. We consider these in turn.

Reinforcement of Existing Practice Patterns

While diffusion of information can create patterns of variable medical care use, legal rules also enforce them. Most notably, medical malpractice law commonly uses standards of negligence that virtually required individual physicians to conform to local standards until quite recently. Under these rules, "local patterns of practice" are the standard against which individual cases are judged to determine negligence. Expert witness testimony in such cases may refer only to local practice standards, not "national" standards, and doctors are (in concept) deemed liable only if they do not meet local standards. Recent case law has modified these rules so that a national norm prevails in some areas of specialty care, but local standards still prevail in the practice of less specialized care. Thus, even though local custom may differ greatly from region to region, legal rules emphasize adherence to local custom as a defense against malpractice suits. (See King, 1986, for discussion and citations).

This issue cannot fully explain the continuance of patterns of variation, since similar patterns occur in countries with greatly differing medical malpractice laws. However, in the U.S. context, they surely serve to reinforce whatever else helps to sustain local practice patterns, even if quite different from other "local patterns" in other regions.

Production and Diffusion of Knowledge

The incentives for producing and disseminating knowledge about medical care differ greatly for two distinct components of treatment—"devices" and
"strategies." For manufactured inputs into the process of producing "treatments," essentially normal market and legal incentives apply. This includes capital and equipment used by medical care providers, and drugs and devices used by the end consumer. In these markets, manufacturers confront standard incentives to produce and disseminate knowledge about the efficacy of their products, since the property right to their product guarantees that they will capture much if not all of the economic gain from demonstrating the value of their product. (An important exception to this statement arises from the Food and Drug Administration's regulation of drugs and medical devices. There, the 17-year horizon on patents begins the moment the manufacturer files the patent, but the lag from that point to actual sales averages over 10 years because of the time required for FDA-mandated testing and other aspects of the regulatory process. Thus, much of the economic value of the patent has become eroded by the time the manufacturer actually begins to sell the product.)

The legal liability of manufacturers also provides an incentive to gather information. If the use of a drug or device harms a patient, the manufacturer can be held liable for those damages under normal rules of law. This provides the usual incentives for manufacturers to test the safety of the product extensively, although FDA regulations about testing make it difficult to know just how much testing would occur strictly from the consequences of liability law.

In the treatment of illnesses, one set of "inputs" greatly differ from specific drugs or devices. These can be characterized as a portfolio of treatment strategies and procedures, the selection and application of which forms a very large part of the practice of medicine. These strategies are often quite formalized, such as "taking a history" and "routine physical examination," designed to assist in the process of diagnosis. Beyond these generic strategies, however, doctors must solve often-complex diagnostic problems with a myriad of alternative tests available, each with imperfect accuracy, and must then select (conditional on a diagnosis) among alternative treatments. Sometimes the doctor will perform these treatments directly, and sometimes will refer the patient to a different doctor who specializes in that type of treatment (perhaps for surgery)

7I use the term "strategy" to connote a particular way of assembling previously available activities. For example, a strategy might consist of the decision to use a laboratory test, possibly followed by other tests, conditional on the output of the first test. At some point, the strategy involves a decision to treat, referral to a specialist, or doing nothing. Treatment might consist of decisions among several drugs that might be used, and way to initiate and then adjust the dosage, according to the patient's response. Similarly, a strategy might include a combination of drugs and a sequence of radiation treatments for a patient with a particular type of cancer.

I use the term "procedure" to connote an activity that is discrete and involves physical activity by a doctor, like a surgical intervention. Research in surgical procedures creates new ways to do surgery on a particular medical problem. Recent changes in the tools available to surgeons has dramatically changed both the procedures and strategies used by surgeons for some illnesses or injuries. For example, arthroscopic surgery allows correction of a number of defects of joints that produce much less "collateral damage" than previous techniques. Many professional athletes have had such surgery, often returning to play within weeks of having had the surgery. Under older methods, knee surgery often took an athlete out of action for months, and sometimes years.
of a particular type). The strategies for such diagnosis and treatment choices form the crux of most "applied" medical knowledge, and the decisions that arise from the application of such strategies lead to the variations in regional patterns of care described earlier.

Economic and legal incentives differ greatly regarding "strategies" from those applying to drugs, devices, equipment, and so on. There are no property rights to strategies, and no liability attached to their use except in the particular instance of a single doctor-patient combination. Hence, the incentives to test the safety and efficacy of a strategy differ greatly from those for drugs and devices.

The role of property rights and product quality has been discussed extensively in the economics literature. The standard argument says that when the manufacturer can capture the profits of higher quality, then investment in quality will proceed appropriately until the marginal cost of quality equates with the marginal social benefit. Further, within that broad rubric, manufacturers have strong incentive to be able to document the quality of their product to buyers, and comparison shopping enhances product quality (Dionne, 1984). However, no property rights exist to the use of strategies; their production is a public good.

The production and dissemination of strategies takes place in medical journals, just as does the production and dissemination of ideas in economics. Indeed, the incentives for their production are quite similar, and take place under similar conditions. Many medical journal articles describe treatment strategies and argue why they should be successful. Doctors who invent and test such strategies typically work in medical school settings, where their promotion depends at least in part on their success in publishing.8

An analog exists to the work commonly undertaken by economists, namely studying the accuracy of various economic models. Most economic research is published in the public domain, and the primary benefit to the researcher comes in the form of personal prestige and career advancement associated with such publications. Some economic research is privatized, like economic forecasting for specific industries or firms. The rewards to doing this depend on the importance of economic conditions to the firm in question and the accuracy of the work, but clearly much of the gain from such research is private, and can be rewarded accordingly. The scope of the firm determines how much of such analysis is undertaken; General Motors and AT&T certainly conduct much more of this type of work than do Mom and Pop's Ice Cream Store or even Ben and Jerry's Ice Cream.

8Physicians working in medical schools are often part of large group practices, with the income from the medical practice pooled and used to pay the salaries of members of their department. The analog for economists would be that a major portion of each faculty members' time were spent in consulting activities, the income from which went to a department-wide pool. A substantial fraction of the total department salaries comes from such pools in medical schools, rather than as guaranteed support from the Dean, as is the case in colleges of arts and science.
Finally, some economic research can be highly privatized with immense
current rewards for devising a successful strategy. Research into the behavior of
prices in the stock market presents an obvious area. The occasional "retire-
ment" from academia into the private sector of previous practitioners of the
random walk model suggests that publication of results regarding strategies in
the stock market will have a distinctly filtered flavor, with those studies support-
ing random walk findings remaining in the public domain, and those finding
chunks in the random walk remaining private.

In medicine, however, almost no comparable mechanisms exist to exploit
the development of a novel and successful strategy for treatment of patients;
hence, almost all of the research remains in the public domain. It seems worth
considering just why this happens in medicine.

The problem of devising successful strategies for carrying out complex and
often-repeated tasks has been resolved on other service-sector industries
through the vehicle of franchising. Fast-food restaurants, franchised auto repair
facilities, some educational activities (such as tutoring for college and graduate
school entrance examinations), and income tax services provide examples of
how organizations can create standard methods of accomplishing certain tasks,
require their use in local franchises, and carry out research to improve the
practices. In each of these cases, fairly precise "procedure manuals" describe
the knowledge base accumulated by the parent organization, and standardiza-
tion of the product emerges through use of such manuals and repeated
inspection of local operations.

However, franchising is not likely to succeed in medicine because of the
extended complexity of problems confronting most practitioners. The tasks are
too complex to specify in detail, and hence are not amenable to "standard"
solutions that can be written down. Basic practice guidelines have been pub-
lished in several settings, like the standard "treatment guides" that most young
doctors (and many experienced doctors) carry around in the pockets of their
white coats. These provide "first steps" in the diagnosis and treatment of
diseases that doctors can look up rapidly. The degree to which they are used
varies greatly from doctor to doctor, and to my knowledge, few if any medical
treatment organizations, even large closed-practice HMOs, rely on these as
more than sources of advice to doctors.

As an alternative to standardized procedures, medicine turns to the con-
cept of "the professional," whose judgement becomes the standard of excel-
rence. Arrow's discussion of professionalism bears repeating here—we establish
professions, put professionals in a position of special trust, and then depend
upon their judgement, in part because we have no readily available alternative.
The activities are too complicated and numerous to pre-specify, so the gains
from standardization do not exist.

An additional problem arises from the rate of production of new knowl-
edge within the biomedical sciences. Medical journals proliferate at an astonish-
ing rate, and no single individual can hope to "keep up" even within narrow
realms of medical practice. Most doctors provide medical treatment for a large number, often hundreds of different diseases, and they must be prepared to diagnose an even broader set of diseases. Further, unlike the case of “curing” automobiles, cooking fast food, or completing tax returns, the intrinsic variability of patients’ symptoms and responses makes it difficult to write down a protocol that captures the intricacies of many medical problems. “Protocol” medicine (or in a more pejorative form, “cookbook” medicine) confronts intrinsic limitations because of the complexity of problems confronting physicians. Much of their task in diagnosis involves pattern recognition, one area where humans remain superbly more effective than computers.

Supervising professionals in this type of setting is also quite difficult, because of the idiosyncratic “job shop” nature of each treatment. Even supervision on the basis of aggregate performance (using tools like time per patient, or outcomes) requires complicated adjustment for the mix of patient illnesses treated by each doctor, and within those, the severity of illness. Similar issues arise in the supervision of legal services, architectural services, and so on. For these reasons, most large groups of physicians still have little direct supervision of individual doctors, and almost no physician-firms operate in multiple sites. Even large groups of physicians commonly associated with the same name—such as the Kaiser health maintenance organization activities in northern California, southern California, Oregon, and Hawaii—have completely separate physician groups under contract to direct the medical staffs of those HMOs.

Not only are there few organizational incentives for developing and disseminating information about “strategies” for diagnosis and treatment, but legal considerations also matter. As noted previously, behaving according to local custom forms a defense for practitioners in malpractice suits. But the inventor of a new strategy has no liability regarding its subsequent use. Thus, once a strategy has been invented, the usual product liability constraints on quality do not apply. In contrast, a drug or device manufacturer is commonly a co-defendant in medical malpractice trials, even if the operation of the device or the performance of the drug is not the major question. Inventors of strategies have no legal incentives to provide evidence about their safety or efficacy. This in turn inhibits the development of information about strategies, particularly those arising from the usual path of “academic medicine.”

**Improving the Quality of Medical Information**

Several changes in legal and organizational incentives would lead to increased production and dissemination of medical knowledge. One obvious step

---

9Stigler (1973) offers a whimsical but nevertheless illuminating discussion of how imposition of liability would affect academia.
would alter the *de facto* legal definitions of medical liability to include national, rather than local norms.

Allowing doctors to patent treatment strategies offers a tantalizing step into a market economy where "professionalism" has previously remained. This would be a two-edged sword, however; doctors who produced and patented a strategy for treatment could reap potential profits, but they would also incur liability for subsequent use of that strategy throughout the country. The analog to medical devices here is quite strong; manufacturers have both the profits and the liability associated with the use of their devices. However, patents and other property rights protection have worked poorly in other areas similar to "medical strategies," including legal strategies, art, architectural design, and so on. Several prominent cases involving plots for movies, themes for music, and such show that there can be some legal protection for ideas, but it is intrinsically difficult, and one would be foolish to expect too much improvement in the variability of medical care merely by strengthening property rights to ideas.

In lieu of such changes in the legal system to enhance the normal competitive forces for production and dissemination of information about treatment efficacy, we can expect to see continued widespread variability in the use of medical interventions, and an underproduction of both efforts to produce and use information about treatment efficacy.

In a second-best world, continued efforts to enhance the production and use of such knowledge by government have high payoffs. Phelps and Parente (1990) have estimated the expected returns to information by estimating the welfare loss arising from variability in the use of medical care. Even ignoring the obvious effects that biased rates of use (too much or too little care on average) would produce, they showed that the welfare gains from reduced variability exceed by one to two orders of magnitude the costs of producing and disseminating information about the proper methods of treating illnesses. The federal government has established the Agency for Health Care Policy and Research (AHCPR) to support such studies, but the funding levels remain minuscule compared with those for basic biomedical research. For example, in fiscal year 1991, the AHCPR budget was about $120 million, and the overall budget for the National Institutes of Health was nearly $9 billion, 75 times larger than the AHCPR budget. The apparently high returns to knowledge in this area make further government investment in this public good a high priority goal.

**This work was supported by the Agency for Health Care Policy and Research, Grant No. R01-6366. I wish to thank Carl Shapiro, Joseph Stiglitz, and Timothy Taylor for highly detailed and very helpful comments on a previous draft, and Joseph P. Newhouse and Henry Aaron for their comments and for initiating this paper.**
References


This article has been cited by:


6. Wan S. Jung, Mun-Young Chung, E. Soo Rhee. 2018. The Effects of Attractiveness and Source Expertise on Online Health Sites. *Health Communication* **33**:8, 962-971. [Crossref]


8. David H Howard, John D Roback, David J Murphy. 2018. Trends in transfusion rates after the FOCUS trial. *Journal of Comparative Effectiveness Research* **7**:2, 113-120. [Crossref]

9. Jostein Grytten, Lars Monkerud, Rune Sørensen. Practice Guidelines and Practice Variation: Diagnostic Technology in Maternity Care 505-517. [Crossref]


15. David H. Howard, Yu-Chu Shen. 2014. Trends in PCI Volume after Negative Results from the COURAGE Trial. *Health Services Research* **49**:1, 153-170. [Crossref]


17. Jostein Grytten, Lars Monkerud, Rune Sørensen. Practice Guidelines and Practice Variation: Diagnostic Technology in Maternity Care 1-14. [Crossref]


21. Shahrokh Esfandiari, Reza Majdzadeh, Jocelyne Feine. 2011. Types of Canadian Dentists Who Are More Likely to Provide Dental Implant Treatment. *Implant Dentistry* **20**:1, 76-84. [Crossref]

22. Gang Fang, John M. Brooks, Elizabeth A. Chrischilles. 2010. A New Method to Isolate Local-Area Practice Styles in Prescription Use as the Basis for Instrumental Variables in Comparative Effectiveness Research. *Medical Care* **48**:8, 710-717. [Crossref]


24. Besiki Stvilia, Lorri Mon, Yong Jeong Yi. 2009. A model for online consumer health information quality. *Journal of the American Society for Information Science and Technology* **60**:9, 1781-1791. [Crossref]


34. Weng-Foung Huang, Yi-Wen Tsai, Fei-Yuan Hsiao, Wen-Chun Liu. 2007. Changes of the prescription of hormone therapy in menopausal women: An observational study in Taiwan. *BMC Public Health* **7**:1. [Crossref]

35. Christopher R. Browning, Kathleen A. Cagney. Health, Neighborhood Disadvantage. [Crossref]


42. M. Kate Bundorf, Kevin A. Schulman, Judith A. Stafford, Darrell Gaskin, James G. Jollis, Jose J. Escarce. 2004. Impact of Managed Care on the Treatment, Costs, and Outcomes of Fee-for-Service Medicare Patients with Acute Myocardial Infarction. *Health Services Research* **39**:1, 131-152. [Crossref]

43. Etienne E. Pracht. Health Insurance and Managed Care . [Crossref]


62. Wilhelmina A. Leigh. Will Greater Competition Improve the Markets for Health Care Services and the Health Outcomes for Black Americans? 93-115. [Crossref]

63. Robin C. Sickles, Paul Taubman. Chapter 11 Mortality and morbidity among adults and the elderly 559-643. [Crossref]

64. Elena Granaglia. 1996. The Italian National Health Service and the Challenge of Privatization. *South European Society and Politics* 1:3, 155-171. [Crossref]


66. Charles E. Phelps. 1995. Perspectives in health economics. *Health Economics* 4:5, 335-353. [Crossref]


69. Laurence Baker, Joel C. Cantor. 1993. Physician Satisfaction Under Managed Care. *Health Affairs* 12:suppl 1, 258-270. [Crossref]


71. Joseph E. Potter, Axel I. Mundigo. Fertility Planning 733-759. [Crossref]

72. James B. Rebitzer, Mari Rege, Christopher Shepard. Influence, information overload, and information technology in health care 43-69. [Crossref]