Ana Johnson
and
Thérèse Stukel
Medical Practice Variations
Health Services Research
10.1007/978-1-4899-7573-7_65-1

Small Area Analysis and the Challenge of Practice Variation

John Wennberg[1], Klim McPherson[2, 3] and David Goodman[4, 5, 6, 7]
(1)The Dartmouth Institute for Health Policy and Clinical Practice, 35 Centerra Parkway, Suite 300, 03766 Lebanon, NH, USA

(2)Nuffield Department of Obstetrics and Gynaecology, New College, NDOG, Level 3, Women's Centre, John Radcliffe Hospital, OX3 9DU Oxford, Oxfordshire, UK

(3)University of Oxford, OX1 3BN Oxford, UK

(4)The Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, NH, USA

(5)Department of Pediatrics, Dartmouth Hitchcock Medical Center, Lebanon, NH, USA

(6)Department of Community and Family Medicine, Dartmouth Hitchcock Medical Center, Lebanon, NH, USA

(7)Geisel School of Medicine at Dartmouth, Hanover, NH, USA

John Wennberg (Corresponding author)
Email: john.wennberg@dartmouth.edu

Klim McPherson
Email: klim.mcpherson@new.ox.ac.uk

David Goodman
Email: david.goodman@dartmouth.edu

Abstract

The study of variations has challenged long-standing theories that the utilization of medical services is primarily determined by the incidence of illness, the constraints of medical science, or patient preferences. This chapter traces the development of empirical research in small area variation in the United States, England, and other countries and the concurrent evolution of explanatory theories. Variations in health services are marked across every level of care – small areas, tertiary regions, and health care providers. The majority of this variation is not explained by patient differences. We categorize different sources of variation into (1) effective care, where benefits far exceed harm, (2) preference-sensitive care, where there are a variety of options for patients, each with benefits and trade-offs, and (3) supply-sensitive care, where varying supplies of health care resources lead to differences in the frequency of care, with scant evidence that more services benefit patients. Each of these explanatory categories is linked to specific clinical and policy remedies that, if implemented, would significantly improve health care.

Electronic supplementary material
The online version of this chapter (doi: 10.1007/978-1-4899-7573-7_65-1) contains supplementary material, which is available to authorized users.

Introduction

Throughout most of the twentieth century, and to an important extent today, public policy for health care has been guided by the theory that the utilization of medical services is governed primarily by the incidence of illness, the constraints of medical science, and the preferences of the individual patient. This happy equilibrium is achieved because patients, lacking enough knowledge to make their own clinical decisions, delegate the choice of treatment to their physicians who, by virtue of their clinical training, lifetime commitment to learning, and acquired skills in diagnosing patient preferences, are able to make the best recommendations for the individual patient. Although asking physicians to decide what patients need opens up the opportunity for physicians to induce demand, potential conflicts of interest are held in check by medical ethics, aided by professional utilization review and malpractice lawyers. Thus, the clinical decisions made by individual physicians, acting as agents for their patients, sum up to the "right intervention rate": the amount of care required to meet the needs and preferences of the population.

Public policy governing resource investments has also assumed that physicians in their clinical agency role can also serve as stewards of society's resources. When supply is in excess of demand, hospital beds and physician services would be underutilized and the "market" cleared of excess capacity. But when supply is inadequate, for example, when hospitals are fully occupied or physician appointments are difficult to get, health care is rationed until society steps in to increase capacity. Under the agency paradigm, it is the duty of a compassionate society to make sure there are enough hospital beds, physicians, and nurses to assure that utilization is not constrained for lack of capacity.

Thomas Kuhn states that when evidence is seriously at odds with the predictions of established theory, the underlying scientific paradigm can be overthrown, replaced by new theory (Kuhn *2012*). While Kuhn's analysis centered on unexplained anomalies in the physical sciences, the discovery of unwarranted geographic variation in health-care delivery – which is defined as variation that clearly isn't explained by illness, medical science, or patient preference – has increasingly come to undermine clinical agency theory as an adequate explanation of the behavior of the health-care system. That is, there is no strong and dependable link between patient needs and preferences and physicians' medical care recommendations. This chapter will describe the evidence from medical practice variation that undermines the dominance of such a theory. Once that is achieved, a coherent policy for health-care reform emerges.

The first, and arguably the most dramatic, assault on clinical agency theory is J. Alison Glover's study in the 1930s, in which he uncovered what he labeled the "strange bare facts of incidence": the striking variation in tonsillectomy rates among British school districts, without any apparent relationship to illness or the socioeconomic factors that condition a patient's behavior in seeking medical care (Glover *1938*). Glover linked geographic variations to wide differences in medical opinion about the value of tonsillectomy among the school physicians responsible for referring children for surgery. His strongest argument for the importance of medical opinion in determining the rate of tonsillectomy was based on a natural experiment in which he monitored the changes in tonsillectomy rates that followed

a change in medical personnel – the school health officer responsible for diagnosis and referral of children for surgery. His most famous case was that of Dr. Garrow, who replaced an unnamed health officer in the Hornsey Borough school district. Following the change, tonsillectomy rates dropped about 90 % (see Fig. 1).

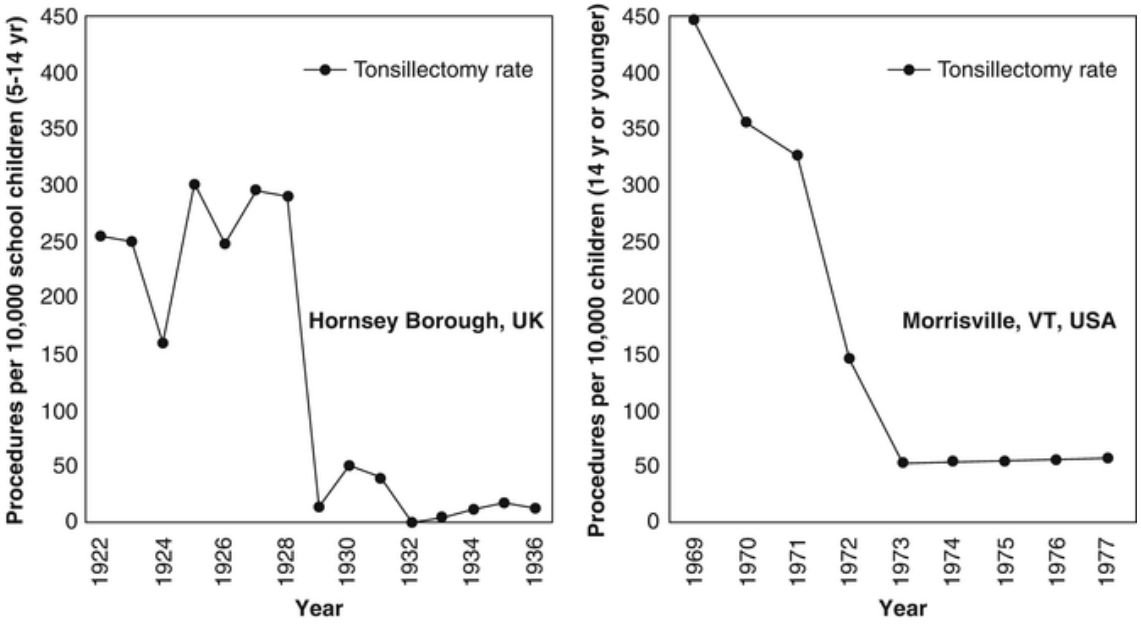


Fig. 1

The tonsillectomy rate per 10,000 children in the Hornsey Borough school district (1922 through 1936) and in the Morrisville, Vermont, hospital service area (1969 through 1977). *Left*, tonsillectomy rates before and after Dr. Garrow's appointment as school physician in 1929 (Adapted from Glover, J. Alison. The incidence of tonsillectomy in school children. *Proceedings of the Royal Society of Medicine* 1938;31:1219–1236. Reprinted in the *International Journal of Epidemiology*, 2008;37:9–19. Reproduced by permission of the Royal Society of Medicine Press, London, and Oxford University Press, Oxford, UK). *Right*, changes in tonsillectomy rates associated with feedback and review (Data extracted from Table III in Wennberg JE, Bowers L, Parker R, and Gittelsohn AM. Changes in tonsillectomy rates associated with feedback and review. *Pediatrics*, Vol. 59, Pages 821–826, Copyright 1977 by the American Academy of Pediatrics. Used with permission from *Pediatrics*)

Work in New England by Wennberg and colleagues in the 1970s and 1980s greatly extended the reach of variation studies and the depth of the evidence contradicting clinical agency theory. An annotated bibliography of this work can be accessed on the website of the Dartmouth Institute for Health Policy and Clinical Practice (Wennberg Anthology *2015*). It wasn't just tonsillectomy, nor was it just elective surgery that varied. By the 1970s, practice variation had been reported for a number of common surgical procedures (Lembcke *1952*; Lewis *1969*; Dyck et al. *1977*; Vayda *1973*). With a few important exceptions, such as hip fractures or heart attacks, the population-based rates of delivery for most medical services and conditions were highly variable among regions, more so than hysterectomy (usually a highly discretionary procedure with many possible target outcomes for patients).

Medical resources – the per capita number of hospital beds, nursing home beds, primary care physicians, surgeons, and medical specialists – also varied substantially and were correlated with variation in service use.

Beginning in the 1990s, the Dartmouth Atlas Project extended the New England analysis to the entire US Medicare population, resulting in information on practice variation for some 3,400 hospital service areas and 306 hospital referral regions (http://www.dartmouthatlas.org/). The Dartmouth

work has led to an analytic framework (*Wennberg et al.*) for studying unwarranted variation that distinguishes three categories of care useful for differentiating the mechanisms that supply plays in influencing utilization:

Effective care: Evidence-based interventions for which the benefits exceed the harm (and for which, by professional consensus, there is no "reasonable" alternative option). Examples include hospitalization for hip fracture, childhood immunizations, and beta-blockers after heart attack. Unwarranted variation in the use of such treatments among eligible patients reflects failure to deliver appropriate care.

Preference-sensitive care: Interventions for conditions for which risks and benefits differ across treatment options and patient attitudes toward these outcomes also vary. Examples include knee replacement vs. medical management for arthritis of the knee, PSA screening for prostate cancer vs. watchful waiting, hysterectomy for heavy menstrual bleeding rather than endometrial ablation, and "high-tech" end-of-life care vs. hospice care. To the degree that variation in the use of preference-sensitive care reflects physician opinion that diverges from what informed patients want, it is unwarranted.

Supply-sensitive care: Everyday services where the supply of resources has a major influence on the frequency of use, such as physician visits, referrals to specialists, hospitalization for medical conditions, stays in ICUs, diagnostic tests, and imaging exams. In the absence of evidence that greater intensity of use results in better outcomes, the variation should be viewed as unwarranted.

In this introductory chapter, a summary is provided of the New England/Dartmouth Atlas studies of practice variation, with emphasis on those studies that have influenced the understanding of the role of supply in influencing utilization. This is followed by brief comments on recent international studies of medical practice variation, the findings of which are consistent with, and indeed predicted by, the evidence accumulated by the Dartmouth Atlas Project. The problem of unwarranted variation seems to be ubiquitous and in need of an international effort to remedy it.

Finally, the use of the Dartmouth analytic framework for identifying strategies to address unwarranted variations, concentrating on supply-sensitive and preference-sensitive care, is illustrated. The strategies include conducting outcomes research to improve the scientific basis for clinical decision-making; establishing informed patient choice as the ethical basis for making preference-sensitive treatment decisions; rationalizing the frequency of use of supply-sensitive care by building and testing clinical pathways for managing care over time, particularly for those with chronic illness; encouraging the reallocation of excess capacity in acute care hospital resources and physician specialists (as now appear to exist in many US hospital referral regions); and establishing population-based data systems to promote transparency and monitor performance through routine health-care measurement and Atlas projects such as those that are now being used in a number of OECD countries.

The Vermont Variations

The systematic and analytic work in documenting variation in local practice began in the late 1960s in the state of Vermont when Wennberg and Gittelsohn developed a population-based method for measuring health-care performance that they called *small-area analysis of health-care delivery* (SAA). Designed to support the requirements for the regional planning program that Wennberg then headed, SAA captured information on resource input and utilization for inpatient, ambulatory, and nursing home care and supported a detailed analysis of case mix and treatment regimes for

hospitalized patients through a patient-level hospital discharge abstract database covering virtually all Vermont residents. Using Lembke's patient origin approach to defining geographic boundaries of a hospital service area (Lembcke *1952*), each minor civil division in Vermont was assigned to the hospital most often used by its resident population. This resulted in 13 geographically distinct population hospital service areas (HSAs) for which population-based rates were then calculated for resource inputs (such as hospital beds and physician labor) and utilization (such as surgical procedures, hospitalization, nursing home admissions, physician visits, and diagnostic procedures). By virtue of the way the HSAs were defined, the large majority of services in each were delivered by local physicians and occurred at local hospitals, clinics, doctor's offices, and nursing homes. The small-area model thus provided a framework for characterizing local medical variability and testing hypotheses concerning the role of illness, local medical opinion, capacity, and clinical science in influencing demand.

The Vermont small-area analysis, initially published in *Science* (Wennberg and Gittelsohn *1973*) , unmasked a surprising picture. While expecting to find a rural health-care system characterized by underservice, the researchers found instead vast and highly significant variations in the deployment of resources and the utilization of services among neighboring communities, without any apparent rhyme or reason. Here is a synopsis of the major findings and interpretations:

Extensive Small-Area Variations in Care Delivery

Virtually all aspects of health-care delivery, including physician services, hospitals, nursing homes, diagnostic tests, and surgery exhibited extensive variation among the 13 Vermont hospital service areas. As an example, hospitalization rates for most causes of admission varied two- to threefold; tonsillectomy rates varied tenfold; hysterectomy varied threefold; X-rays, electrocardiograms, and laboratory tests varied four-, six-, and sevenfold, respectively; and nursing home admissions varied nearly sixfold.

Illness and Patient Behavior Did Not Explain Variation in Supply and Utilization

Variation in supply and utilization was inversely correlated with need. HSAs with older populations and lower per capita income had fewer physicians per capita; greater spending for hospitals and physician services was uncorrelated with age- and sex-adjusted mortality and perinatal mortality. Direct evidence for the lack of an important role of illness and patient behavior in seeking care to explain variation in utilization was obtained through a household survey (Wennberg and Fowler *1977*). The survey was conducted in six Vermont hospital service areas that differed twofold or more in hospitalizations, surgeries, and spending. It revealed that the residents of these areas were remarkably similar in self-reported illness, insurance coverage, and socioeconomic status. Moreover, individuals in each of these six regions sought care in equal proportions; on an annual basis, about 75 % contacted their physician at least once. The study concluded that the twofold difference in health-care utilization reflected "post-access provider behavior" (what happened after the patients contacted their health-care provider). This varied according to characteristics of the local health-care system, not the characteristics of patients.

Attributes of the Local Physician Supply Influenced Utilization

The specialty mix of physicians correlated with the mix of services provided to a defined population. Residents living in hospital service areas with more surgeons per 10,000 had more surgery at all levels of complexity; areas with more general practitioners doing surgery had higher rates of less complicated surgery; and populations living in hospital service areas with more internists underwent more diagnostic tests and had lower surgery rates.

Flaws in Public Policy Governing Capacity Were Uncovered

The variation in institutional measures of "need" for hospital beds bore no relationship to variation in population-based measures of capacity or utilization. The health-care organizations responsible for certifying the need for hospital beds based their determination in part on the hospital's occupancy rate (which is the percentage of hospital beds that on average are occupied by a patient). This would make sense if health-care demand corresponded to that predicted under clinical agency theory – that it reflects illness, constrained by science and patient preference. But it turned out that hospital occupancy rates were uncorrelated with per capita bed supply or hospital utilization (and illness and access varied little) among Vermont HSAs. Thus, the certification of need protocol called for the construction of more beds in some hospital service areas that already had excess capacity. Similarly paradoxical decisions on resource allocation were observed for the building of surgical facilities.

Physician Opinion as Driver of "Demand" for Surgery Illustrated

As it had for Glover, a natural experiment confirmed the importance of medical opinion in determining the rates of surgery, but in this case it was the change in opinion of local physicians, not a change in medical personnel, that influenced the change in medical practice. Information on the rate of tonsillectomy in each HSA was shared with physicians throughout the state without providing an interpretation of the causes of the variation. Subsequently, the degree of variation declined dramatically, with the Morrisville, Vermont, area showing the greatest decline. Prior to feedback, tonsillectomy rates were high enough that an estimated 65 % of children in Morrisville had undergone the procedure by age 15. Upon learning of the high rate in their area, two Morrisville physicians took it upon themselves to review the indications for the operation and discuss their findings with their colleagues. The rates dropped by nearly 90 %, to be among the lowest in the state (see Fig. 1). In an article published in a medical journal, they described the impact of feedback and the changes in local clinical opinion that led to the drop in tonsillectomy rate (Wennberg et al. *1977*). The data and their story convinced many in the medical community about the importance of physicians in inducing demand – and indicated that feedback on variation could lead to rapid change in medical practice.

Emergence of the Professional Uncertainty Hypothesis

Wennberg and Gittelsohn, in weighing the evidence emerging from the Vermont studies, linked the variations in surgery to seemingly idiosyncratic professional uncertainty about the best way to practice medicine:

[T]he variations are more likely to be associated with differences in beliefs among physicians concerning the indications for, and efficacy of, the procedure [than differences in the incidence of disease].

And uncertainty about best practice was linked to inadequate clinical science:

An important reason for uncertainty is that few prospective studies under controlled circumstances have been performed. Because the outcome of one type of service compared to another (or to none at all) is often not known, the variation in therapeutic and diagnostic procedures observed among different Vermont communities cannot be strictly evaluated.

As a consequence in these circumstances, it became entirely plausible that more health care might both cost more and do more harm on average; therefore, more is not necessarily better:

[G]iven the magnitude of these variations, the possibility of too much medical care and the attendant likelihood of iatrogenic illness is [in the absence of strong evidence] as strong as the possibility of not enough service and unintended morbidity and mortality. (Wennberg and Gittelsohn *1973*)

The Patterns of Practice Variation

As the results of the Vermont small-area analysis became available, attention turned to various other comparisons to confirm and explain the variation phenomena in Maine and Rhode Island, then for Boston and New Haven, and ultimately for the USA as a whole, through the Dartmouth Atlas Project.

Surgical Variation

The variation in Maine proved to be similar to Vermont, but the studies more clearly revealed the idiosyncrasies in the patterns of surgical practice (Wennberg and Gittelsohn *1975*). While the overall rate of surgery was often similar among hospital service areas, there was marked variation for specific procedures. The “mix” of surgical rates was distinctive and persistent over time within each area, a pattern that was termed a “surgical signature” and was high for some procedures, low for others, and near the state average for still others (see Fig. 2). For example, Portland’s signature surgery rate for treating benign prostatic hypertrophy (BPH) was 40 % greater than the state average, varicose vein removal rates were 43 % below average, and hysterectomy rates were at the state average. Surgeons serving Augusta residents had different proclivities toward the use of these same procedures. Their rate for varicose veins was 89 % greater than the state average, while hysterectomies and BPH surgery rates were, respectively, 34 % and 23 % less than the state average. This surgical signature pattern of variation, which persisted over longer periods of time, seemed best explained by the Glover hypothesis of differences in local medical opinion on the value of a given procedure.

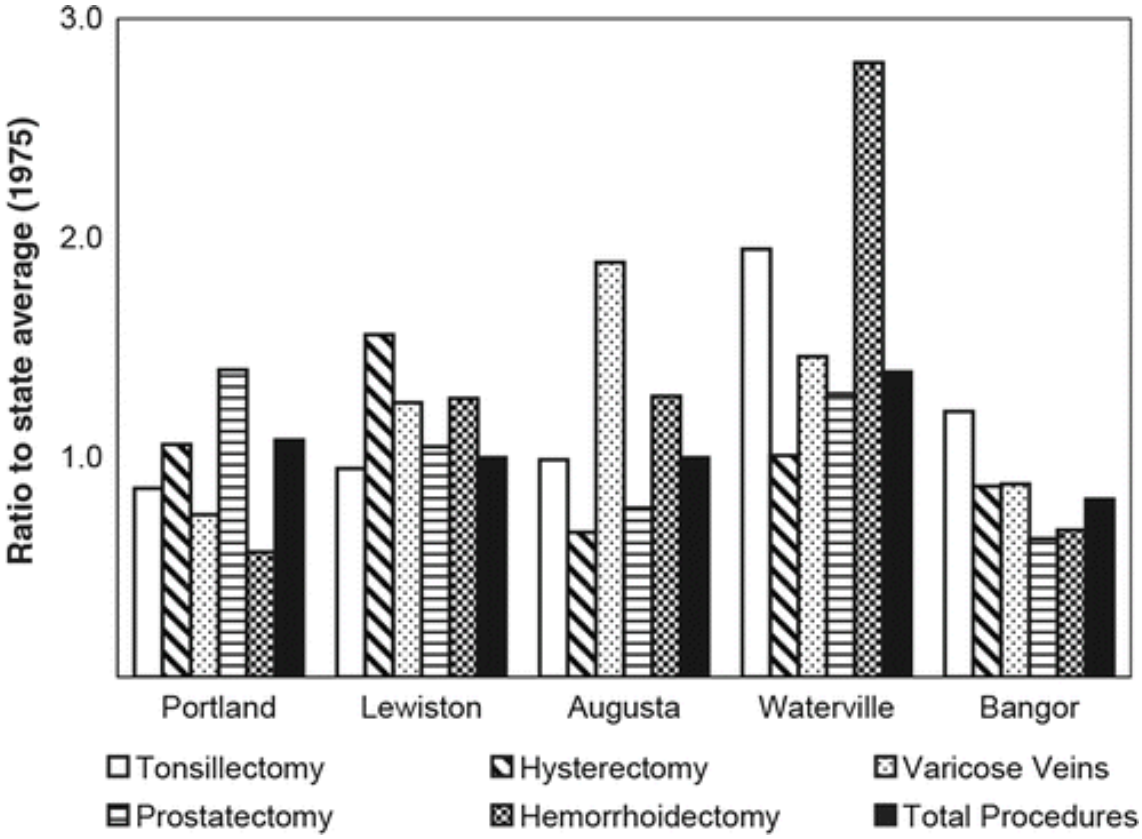


Fig. 2

The surgical signatures of the five most populous hospital service areas in Maine (1975). For each area, the rate relative to the state average for five surgical procedures is displayed (Adapted from Wennberg J and Gittelsohn A. Health Care Delivery in Maine I: Patterns of Use of Common Surgical Procedures. *Journal of the Maine Medical Association* 1975;66:123–130, 149. Used with the permission of the Maine Medical Association)

Medical Variation

In contrast to surgical variation, hospitalization rates for acute and chronic medical conditions tended to be consistent from one cause of admission to another, such that a region with a high rate for congestive heart failure would also have high rates for other chronic conditions such as diabetes, COPD, and cancer. As seen more clearly from the Boston-New Haven studies (see below), regions with high bed capacity experience an across-the-board elevation in admission rates for virtually all medical (nonsurgical) conditions and vice versa. And in contrast to surgical variation, where strongly held medical opinions were behind the variation, the influence of capacity on physician decision-making was subliminal, unrecognized by the physicians who were making the decisions to hospitalize patients.

Low Variation

Not every reason for hospitalizing patients was highly variable. The rates for a few conditions, such as heart attacks, strokes, and bleeding from the gastrointestinal tract, and a few procedures, such as hip

fracture repair, hernia repair, and removal of the colon for cancer, varied relatively little across areas. These “exceptions to the rule” – which together account for less than 20 % of hospitalizations – seemed to follow a pattern of variation predicted by clinical agency theory where illness and medical science drive utilization. A fractured hip, for example, is reliably and easily diagnosed, and there is consensus among physicians that virtually all patients with hip fractures need to be hospitalized. The hospitalization rates for these procedures and conditions serve as benchmarks for medical care where the underlying illness rate is the likely source of variation.

Consistency in the Pattern of Surgical Variation

The degree of variation in surgery appeared to be a property of the procedure (Wennberg and Gittelsohn *1982*). The hernia operation showed relatively little variation among hospital service areas in each state, while tonsillectomies varied a lot, and others were in between (see Fig. 3). The evidence for a consistent ranking in relative variation allowed the development of a scale for classifying hospitalization and surgery according to their relative variation. Procedures or causes of admission with a degree of variation ranked between inguinal hernia and gallbladder surgery were “low variation,” “moderate variation” ranked between gallbladder surgery and hysterectomy, and procedures and causes of admission that were more variable than hysterectomy were “high variation.”

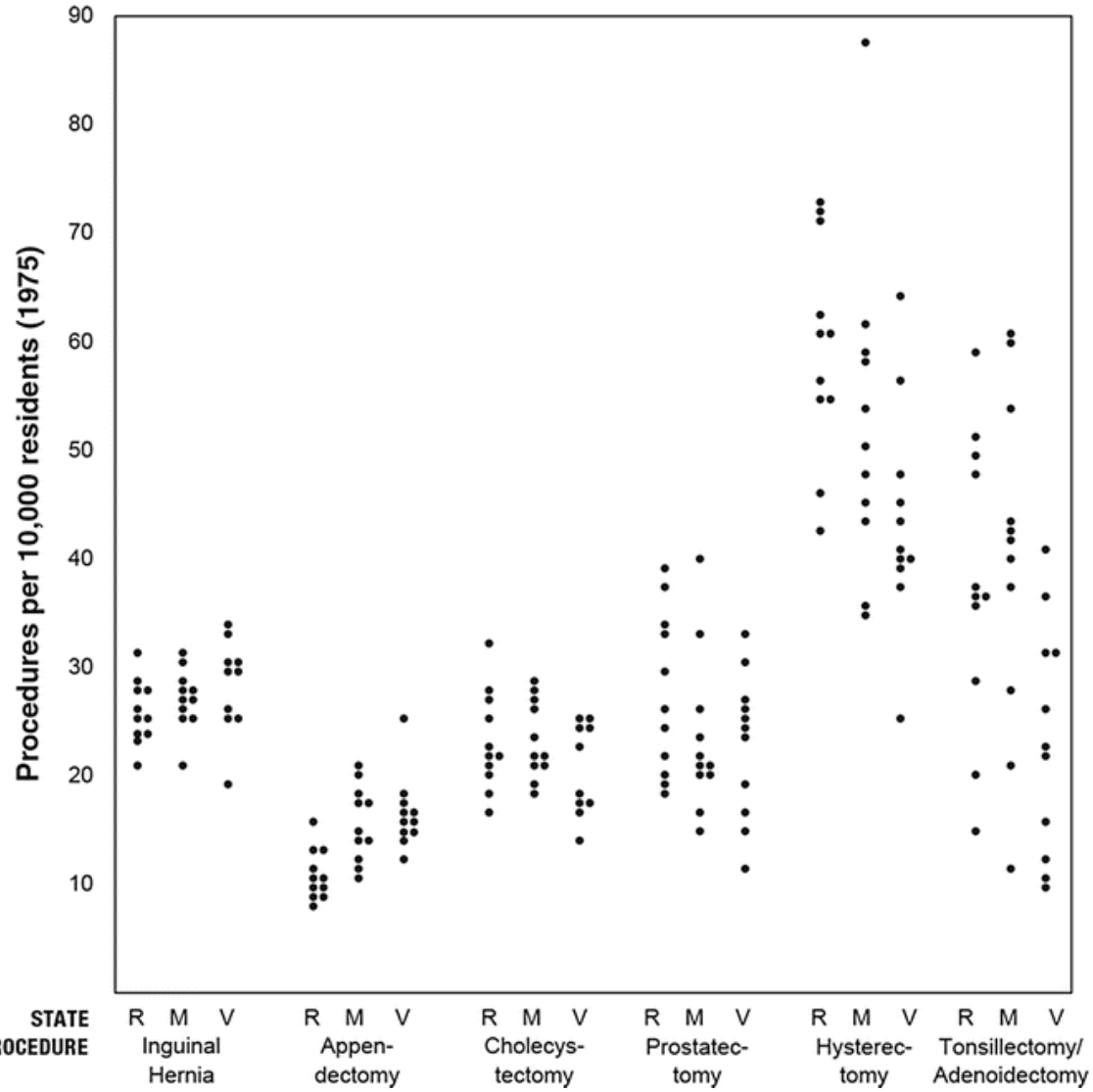


Fig. 3

The surgery rates for six common procedures among the 11 most populated hospital service areas in Rhode Island, Maine, and Vermont (1975). The procedures show increasing variation from *left* to *right*. Each *dot* represents a hospital service area: *R* Rhode Island, *M* Maine, *V* Vermont (Adapted from Wennberg J and Gittelsohn A. Variations in Medical Care among Small Areas. *Scientific American* 1982;246(4):120–134. Used with permission from Scientific American)

How Much of Hospital Care Is "High Variation"?

What proportion of hospital care follows the low-variation pattern (thus conforming to the appropriate pattern predicted by clinical agency theory), and what follows the high-variation pattern (thus conforming to the pattern where medical opinion and capacity strongly influence demand)? To find out, the then newly developed DRG system provided a tool for grouping the myriad number of individual ICD diagnosis codes into meaningful clinical groups. McPherson and colleagues developed a means for measuring variation that distinguished between random and systematic variations, termed

the systematic coefficient of variation (SCV), that allowed for the study of variation among regions with different sized populations and procedures with different mean rates (see below).

Among hospital service areas in Maine, the high-variation utilization profile proved to be ubiquitous (see Table 1); 80 % of hospitalizations for surgery and medical causes of admission were more variable than hysterectomy. Low and moderate variation were the exceptions: only three medical conditions and six operations were less variable than hysterectomy (Wennberg et al. *1984*).**Table 1 Medical and surgical causes of admissions, ranked in ascending order of variation in incidence of hospitalization among 30 Maine hospital market areas, 1980–1982*** (Adapted from Wennberg, J.E., McPherson, K., and Caper P. Will Payment Based Upon Diagnosis-Related Groups Control Hospital Costs? *New England Journal of Medicine* 1984;311:295–300. Used with permission from *New England Journal of Medicine*)

Medical causes of admission	Surgical causes of admission
Low variation	**Low variation**
None	Inguinal- and femoral-hernia repair
Moderate variation	Hip repair except joint replacement
AMI	**Moderate variation**
Gastrointestinal hemorrhage	Appendicitis with appendectomy
Specific cerebrovascular disorders	Major small- and large-bowel surgery
High variation	Gallbladder disease with cholecystectomy
Nutritional and metabolic diseases	Adult hernia repairs except inguinal and femoral
Syncope and collapse	**High variation**
Respiratory neoplasms	Hysterectomy
Cellulitis	Major cardiovascular operations
Urinary-tract stones	Pediatric hernia operations
Cardiac arrhythmias	Hand operations except ganglion
Miscellaneous injuries to extremities	Foot operations
Angina pectoris	Lens operations
Toxic effects of drugs	Major joint operations
Psychosis	Stomach, esophageal, and duodenal operations
Heart failure and shock	Anal operations
Seizures and headaches	Female-reproductive-system reconstructive operations
Adult simple pneumonias	Back and neck operations
Respiratory signs and symptoms	Soft-tissue operations
Depressive neurosis	**Very high variation**
Medical back problems	Knee operations
Digestive malignancy	Transurethral operations
Gastrointestinal obstruction	Uterus and adnexa operations
Adult gastroenteritis	Extraocular operations
Peripheral vascular disorders	Misc. ear, nose, and throat operations

Red-blood-cell disorders	Breast biopsy and local excision for nonmalignancy
Adult diabetes	D&C, conization except for malignancy
Circulatory disorders except for AMI, with cardiac cath.	T&A operations except tonsillectomy
Very high variation	Tonsillectomy
Deep-vein thrombophlebitis	Female laparoscopic operations except for sterilization
Adult bronchitis and asthma	Dental extractions and restorations
Organic mental syndromes	Laparoscopic tubal interruptions
Chest pain	Tubal interruption for nonmalignancy
Transient ischemic attacks	
Kidney and urinary-tract infections	
Acute adjustment reaction	
Minor skin disorders	
Trauma to skin, subcutaneous tissue, and breast	
Chronic obstructive lung disease	
Hypertension	
Adult otitis media and URI	
Peptic ulcer	
Disorders of the biliary tract	
Pediatric gastroenteritis	
Pediatric bronchitis and asthma	
Atherosclerosis	
Pediatric otitis media and URI	
Pediatric pneumonia	
Chemotherapy	

*Modified DRG classification; residual modified DRGs are excluded. See text for explanation of scale of variation. AMI denotes acute myocardial infarction, URI upper respiratory infection, Misc. miscellaneous, D&C dilation and curettage, T&A tonsils and adenoids, and cath. catheterization

The Boston-New Haven Studies

Comparing practice style between Boston and New Haven was intended to test the hypothesis that unwarranted variation is a central feature of even the most advanced health-care market and that here also the explanatory power of the clinical agency theory was inadequate. Previous studies of variation could be discounted on the basis that they did not reflect what was happening in prestigious medical centers such as Harvard or Yale. The first inkling of the magnitude of the differences in medical care use among regions served by academic medical centers came in 1979. In that year, on a per capita basis, expenditures for hospital care for residents of Boston were twice as much as those for residents of New Haven. How could it be that the highest quality care produced for Bostonians could require twice as many resources as the highest quality care for New Havenites?

Differences in Resource Use

In 1982, the per capita number of hospital beds allocated to Bostonians exceeded the New Haven amount by 55 %: 4.5 vs. 2.9 beds per 1,000 (Wennberg et al. *1987*). Physicians serving Boston used 739 more beds than they would have used had the practice patterns of New Haven prevailed. For what clinical purposes were the "extra" beds of Boston used? The completion of the Maine DRG study (described above) provided a method for characterizing in detail the hospitalized case mix of defined populations, both surgical and medical. Only 4 % of the excess beds went to low-variation medical conditions, and this was primarily because of Boston's longer length of stay. Most of the extra beds – 68 % – were used to care for patients with high-variation adult medical conditions. For these conditions it was admission rates – the decision to hospitalize, not how long to keep a patient in the hospital – that were strongly associated with utilization. It was 56 % higher, while the length of stay was 11 % longer.

For minor surgery, the Boston admission rate was 38 % higher and the length of stay 17 % longer – accounting for 12 % of the excess beds. For major surgery, the overall admission rates were nearly the same in the two communities; 12 % of the excess beds associated with these procedures were explained entirely on the basis of Boston's longer length of stay.

The Medical Signature of Boston and New Haven

The extra beds of Boston appeared to have a consistent effect on admission rates, affecting most high-variation causes of admission (see Fig. 4). But the physicians who practiced in Boston or New Haven were unaware that they or their colleagues in the other place practiced differently. In conversations with them, including some who had practiced in both places, it was clear that the effect of capacity on clinical decision-making was subliminal, affecting the admission rate for a host of acute and chronic "high-variation" medical conditions without awareness that this was happening among those responsible for decisions to hospitalize patients. These conversations also made it clear that there was no sense of lack of resources or fear of health-care rationing among New Haven physicians, even though, from a regional as well as a national perspective, the allocation of resources to New Haven residents was modest. The occupancy rate of the hospitals in New Haven, which at that time were relatively low and about the same for Boston teaching hospitals, stood as additional evidence against rationing.

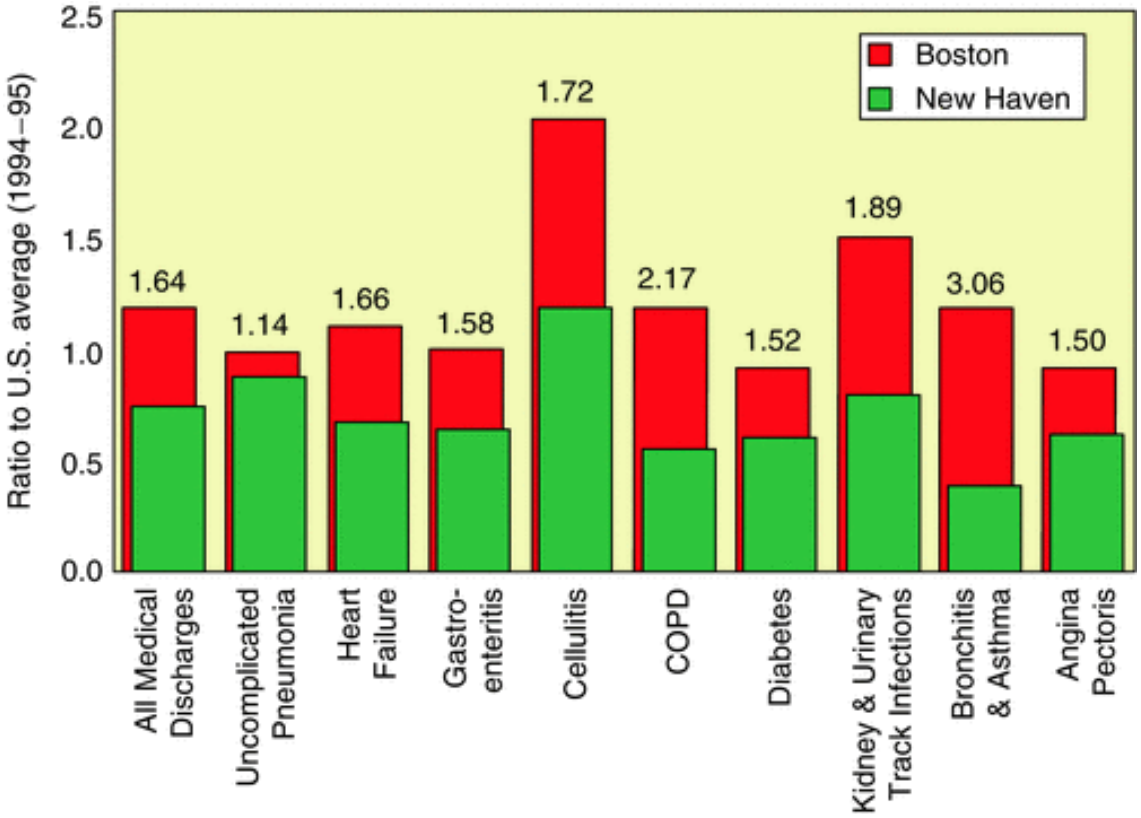


Fig. 4
The medical signature of Boston compared to New Haven (1994–1995). The ratios of discharge rates are given in *black*. Rates are consistently higher in Boston for all listed causes of admission (Adapted from the Dartmouth Atlas of Health Care *1998*)

The Surgical Signature of Boston and New Haven

By contrast, major surgery rates were unrelated to variation in per capita bed capacity, as the procedures followed the idiosyncratic “surgical signature” profile first seen in Maine (see Fig. 5). There was no belief prior to this research about which community had the higher rate, but once the variation seen in Fig. 5 was exposed, the clinicians came up with strong hypotheses based on their understanding of local medical culture. The hypothesis advanced to explain the low rate of carotid artery surgery in New Haven was fascinating. Everyone there seemed to know that Yale neurologists strongly believed that aspirin was a better alternative for reducing the risk of stroke than was carotid artery surgery, and their strong point of view on the matter seemed to keep the surgery rate low.

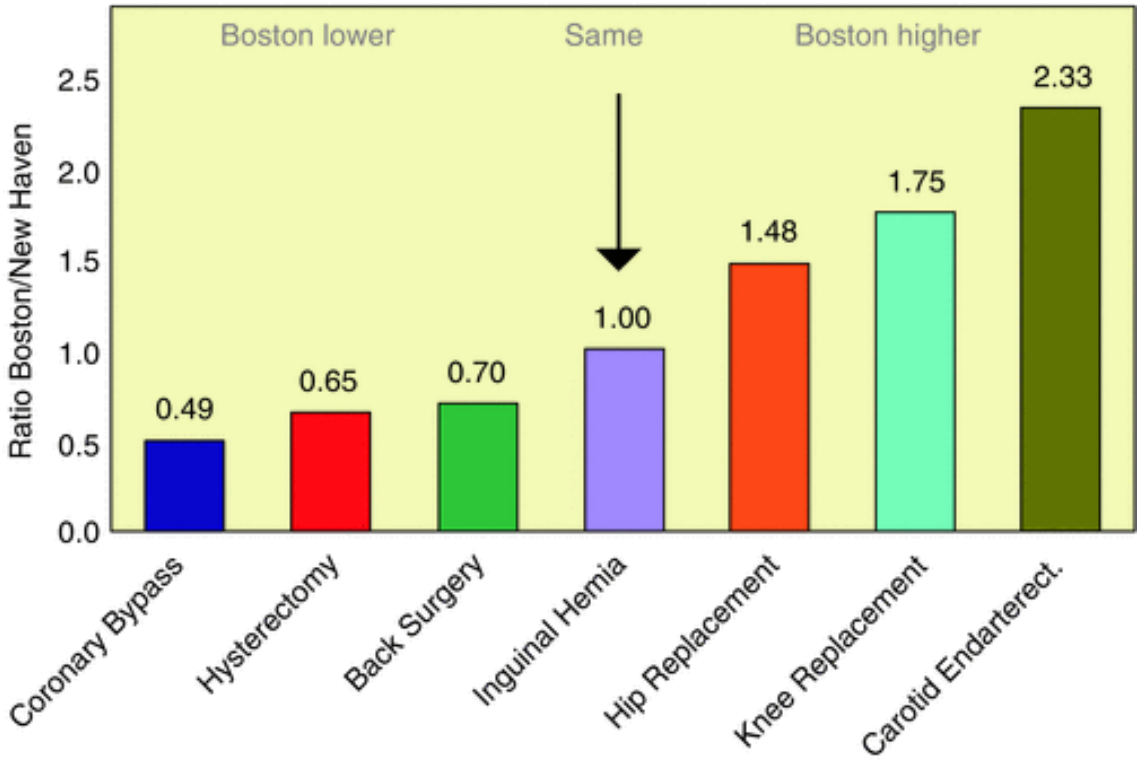


Fig. 5
The surgical signature of Boston compared to New Haven (1994–1995). The ratios of discharge rates in Boston compared to Boston are given in *black* (Adapted from Wennberg JE, Freeman JL, and Culp WJ. Are Hospital Services Rationed in New Haven or Over-Utilised in Boston? *Lancet* 1987;1(8543):1185–1188. Used with permission from Lancet)

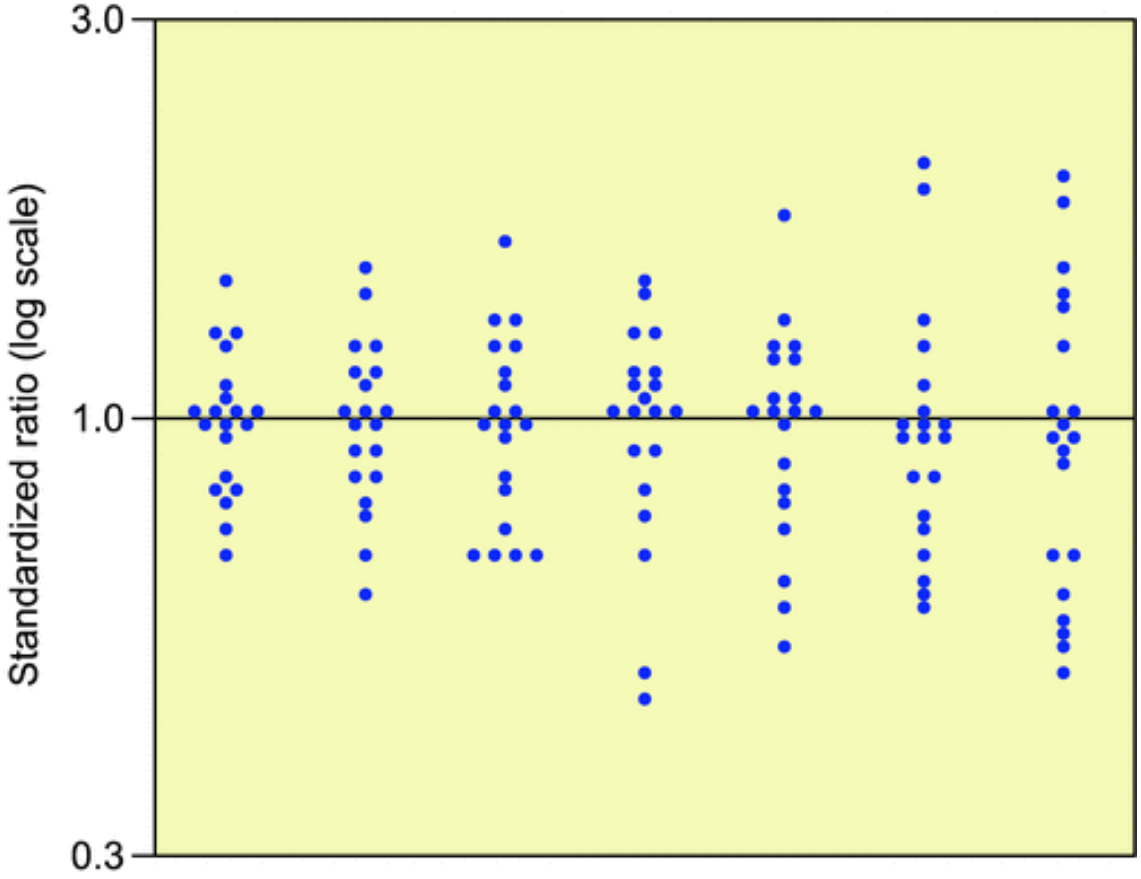
International Variation

Variation Between Countries

Nearly 50 years ago, John Bunker quantified the differences in discretionary surgical rates between the USA and UK (Bunker *1970*). There were surprisingly large differences, not readily explicable by any plausible intrinsic difference in population need between the countries. This raised the obvious questions about both their causes and the consequences. Were the higher rates in the USA evidence of overuse in the USA or underuse in the UK? For Bunker it seemed that the obvious explanation lay in differences in manpower density; there were more surgeons per capita in the USA. But the main question about consequences remained unanswered – were higher rates beneficial or harmful to patient populations?

Since then, international differences between countries have been studied more extensively among OECD countries. Studies of 1989 data by McPherson (*1989*), 2004 data by Scott and colleagues (*2008*), and 2008 data by McPherson et al. (*2013*) all described variation, the extent of which depended on the procedure. It is clear that the highest variations between countries are for back and cardiac procedures, largely between North America and the UK, possibly reflecting the constrained NHS relative to unrestricted markets of an insurance-based system. It is also of interest that the

relative variation of procedures among OECD countries seems to correspond to that predicted by hospital service area variation in the USA, at least for some procedures. For example, in 2008, rates for prostatectomy, hip replacement, and hysterectomy among 17 OECD countries illustrated the high-variation profile and appendectomy the moderate profile, corresponding roughly to the same pattern of variation seen in New England in the early 1970s (see Fig. 6).



	Append-ectomy-males	Append-ectomy-females	C-section	Hip replace-ment-females	Hip replace-ment-males	Hyster-ectomy	Prostate-ctomy
Extremal ratio	2.09	2.47	2.39	3.22	3.34	3.49	3.87
Interquartile ratio	1.29	1.34	1.44	1.29	1.44	1.35	1.87
Coefficient of variation	19.6	22.4	25.4	26.4	28.5	37.9	41.5

Fig. 6

Variation among 17 OECD countries for selected procedures (2008). Appendectomy rates exhibit the least variation and prostatectomy the most. The relative variation of the procedures in the figure is similar to that seen in Maine, Rhode Island, and Vermont in the 1970s (Source: author reanalysis of data reported in McPherson et al. (*2013*))

The variation in hysterectomy rate at the country level showed signs of international convergence between the 1989 and the 2004 studies. However, appendectomy was the only procedure to

demonstrate a clear convergence (see Fig. 7). (Looking at Fig. 7, it is almost as if the "right rate" is around 120/100,000 per annum.) Most interestingly, appendectomy rates became more similar because of declining rates in German-speaking countries. Why? It is suspected that the rates in German-speaking countries were elevated because of particular medical training requirements, as reported in a 1971 article in Medical Care (Lichner and Pflanz *1971*). The findings of this article were widely reported in the lay press. This speaks again to the possibility that information on variation can itself affect medical behavior. Were these international rates to reflect different amounts of morbidity or vastly different preferences, the rate differences would probably have remained with the passing of time.

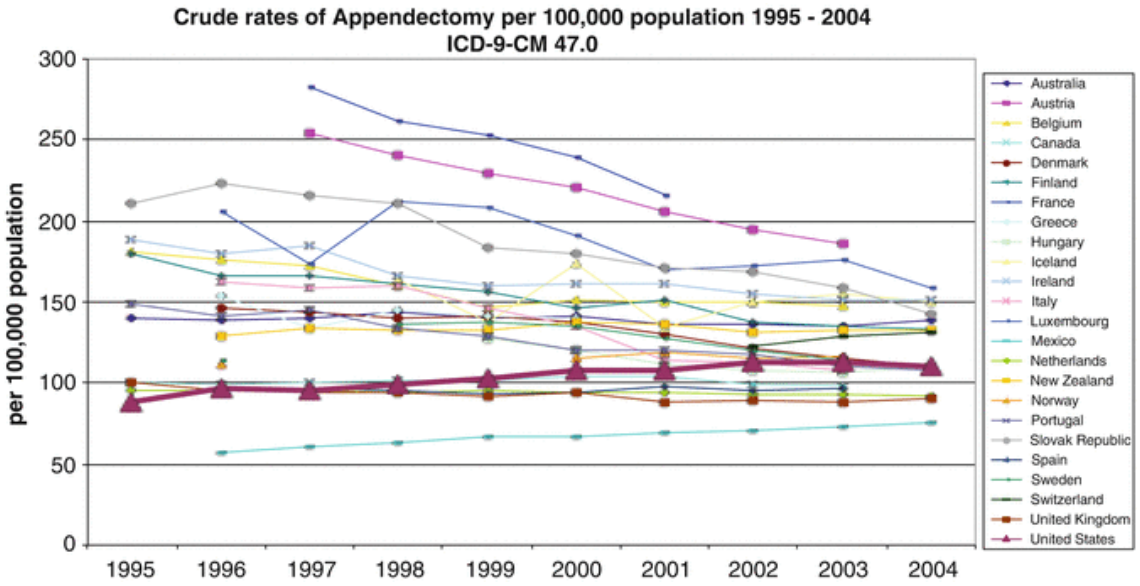


Fig. 7
Crude rates of appendectomy 1995–2004 in 25 OECD countries (Source: McPherson et al. *1996*)

The most recent international contribution to the study of variation is the 13-nation OECD study published in 2014 (OECD Report *2014*). This work analyzes both the differences between small administrative areas within countries and differences between countries. The OECD report shows that cardiac procedures display the highest levels of geographic variation – threefold across countries – and have the highest level of within-country variation for more than half of the countries studied. Knee replacement unsurprisingly varies more than fourfold across countries. Hospital medical admissions, a measure of supply-sensitive care, are more than twice as high in Israel, Germany, and Australia than in Canada.

Small-Area Variation Within Countries

The evidence from New England that a given surgical procedure had a characteristic pattern of regional variation, often driven by clinical uncertainty, led McPherson and Wennberg to ask if the same amount of intrinsic variation on a procedure-specific basis would be found between small geographic areas in other countries. The argument was that, notwithstanding large variations for some procedures between countries, the procedure and not the country would determine the amount of intrinsic variations within countries. If so, this would provide further support for the importance of clinical decision-making as a driver of variation. Physicians everywhere face the same limitations in knowledge concerning the outcomes of care and the preferences of patients, even though the systems

in which they practice may be very different. This was tested by looking at hospital service areas in New England, the USA, NHS regions in the West Midlands, the UK, and counties in Norway (McPherson et al. *1982*). These three countries had quite dissimilar rates for the seven or so procedures examined. For example, hysterectomy was then performed at an annual rate of 540 per 100,000 women in the USA and 118 per 100,000 in Norway.

Testing the hypothesis required a new method for measuring variation, which properly accommodated the random component in estimated annual rates from very different population sizes. Thus, there was a need to measure relative differences in rates as opposed to absolute differences and to subtract the random component of variation that is very dependent on sample size. McPherson and Wennberg called this measure the systematic component of variation (SCV).

The hypothesis turned out to be essentially correct for many common surgical procedures (see Table 2). The patterns of variation observed in the USA with its fee-for-service method of payment and its decentralized "systems" of care were similar to those in Norway and the UK, where physicians are salaried and care is centrally organized. Indeed, the relative variations seen in New England were usually less than in the more centralized health systems in Norway or England, suggesting that these variations are not a product of "uncontrolled" health-care markets. Only in two examples was the SCV measure of variation statistically significantly different. By this measure, hernia repair was more variable among UK regions than it was among counties in Norway and hospital service areas in the USA. A likely reason for this increased variation in the UK was the fact that at the time of the study many patients there were treated nonoperatively – with a truss – an option that was rarely exercised in Norway or the USA. Hysterectomy rates were more variable in Norway, but Norwegian physicians weren't sure why. However, the hysterectomy rate in Norway has recently increased more sharply than it has in England.**Table 2**

Geographic variation in surgery within France, Germany, Italy, Spain, and the UK according to category of care. Variation measured using the coefficient of variation (*CV*) and the systematic coefficient of variation (*SCV*) (Source: data abstracted from OECD Report *2014*)

	France		**Germany**		**Italy**		**Spain**		**The UK**	
Procedure by category of care	CV	SCV	CV	SCV	CV	SCV	CV	SCV	CV	SCV
Effective care										
Hip fracture repair	0.09	0.9	0.11	1.1	0.14	n/a	0.20	4.7	0.16	2.1
Preference-sensitive care										

Hysterectomy	0.18	3.2	0.14	2.3	0.17	n/a	0.21	4.7	n/a	n/a
Knee replacement	0.18	3.6	0.16	2.9	0.20	n/a	0.30	10.7	0.19	3.4
PTCA	0.23	5.2	0.22	4.7	0.23	n/a	0.30	16.4	0.22	n/a
CABG	0.28	7.4	0.24	5.7	0.30	n/a	0.48	23.8	0.30	n/a

In 1995, McPherson and colleagues repeated the small-area variation study in England, this time among the districts of four Regional Health Authorities (McPherson et al. *1996*). The results were very similar to the 1982 small-area study. Here are the SCV measures of variation for each study (Table 3):**Table 3**

Systematic component of variation by procedure between small areas in three countries

Procedure	**The UK**		**Norway**	**The USA**
	1982	**1995**	**1982**	**1982**
Hernia repair	1.7	4.4	0.2	0.6
Appendectomy	2.2	2.9	2.4	1.7
Hysterectomy	3.2	3.7	10.4	4.8
Prostatectomy	4.4	6.2	9.3	5.0
Cholecystectomy	9.1	2.1	1.9	1.7
Tonsillectomy	9.2	18.5	27.5	12.2
Hemorrhoidectomy	11.7	12.2	14.7	12.7

Compared with 1982, the SCV for hernia repair in the UK for 1995 shows a significant increase in variation, possibly because the use of trusses became less common. Cholecystectomy became much more variable, reflecting the availability of a new laparoscopic procedure option that avoided invasive surgery. Tonsillectomy declined in variation, because the procedure was subject to greater scrutiny (Burton *2008*).

Of the 35 procedures looked at in the 1995 study, carotid endarterectomy was the most variable, with an SCV of 56.7, and the highly preference-sensitive procedure, excision of the vas deferens, with an SCV of 55.3. Among a similar number of hospitalizations grouped according to cause of admission,

SCV varied between 49.3 for hypertension and 1.8 for appendicitis. The usual hierarchy between relative ease of diagnosis and consensus on treatment for the effective care category (very low SCV) and almost complete discretion, and hence sensitive to provider opinion, is visible.

Similar work by the King's Fund in 2011 confirms these results more recently (Appleby et al. *2011*). Here they find that SCVs were very similar to those found previously. Tonsillectomy had an SCV of 5.5 in 2005 and 8.4 in 2009. Cholecystectomy had an SCV of 4.5. On the whole, different causes of admission were used in this report, but notably CABG and PCI had an SCV of 8.0 and 14.8, respectively. They concluded that the systematic and routine collation and publication of data on unwarranted variation are the first steps.

With the completion of the large-scale international study of practice variation in 13 OECD countries, data are now available to compare variation among administrative regions within many nations. The pattern of relative variation seen first among hospital service areas in the USA seems to predict well the pattern observed elsewhere. Thus, within France, Germany, Italy, Spain, and the UK, hospitalization for hip fracture (the example of effective care where demand is driven by illness) varies the least in each country. Among the examples of preference-sensitive surgery, hysterectomy and knee replacement vary more, and PTCA and CABG consistently vary the most (see Table 2). Clearly the determinants of the variations demonstrated above are many and complex. The comparisons are generally very consistent; some countries are always high and some always low, suggesting consistent systems or national effect. Health-care systems allow or encourage differentially the capacity for performing different procedures and also preferences among doctors and patients. These will inevitably be culturally determined, around the evidence. The role of different mechanisms for allocating resources is likely to be important and so on. Larger differences in both morbidity and in preferences are of course plausible between countries, but again such large effects seem unlikely. As predicted by studies in the USA, variation in other OECD nations seems to have no rhyme or reason, but clearly must have consequences on health outcomes, and mostly on the cost of care among populations. Indeed, the problem of unwarranted variation is ubiquitous and in need of an international effort to remedy it.

The Challenge of Practice Variation

By extending Glover's analysis of "the strange bare facts of incidence" to show that most surgical procedures and most causes of hospitalization are "high variation," the New England studies greatly increased the challenge of practice variation to the conventional theory that utilization is driven by illness, patient preference, and medical science. In more recent years, small-area analysis has served as a framework for analyzing and interpreting variation uncovered by the Dartmouth Atlas Project and for defining much of the research agenda that Dartmouth faculty and their colleagues have pursued. In this section, the importance of illness, medical opinion, and capacity is further considered as drivers of utilization in the preference-sensitive and supply-sensitive categories of care, and strategies are discussed for reducing unwarranted variation that address fundamental flaws in clinical agency theory.

The Role of Illness Rates

The extent and magnitude of variation uncovered by small-area analysis raised a significant challenge to the assumption that utilization was primarily determined by illness and the socioeconomic

circumstances of patients. It just didn't make clinical or epidemiological sense that illness could vary to coincide with the strange way that surgery varied, as revealed by the surgical signature phenomenon or that illness was behind the apparent influence that bed capacity exercised on the clinicians serving Boston and New Haven, an argument that would require feedback loops linking community morbidity with decisions to construct hospitals and require that illness be higher in Boston for virtually all forms of acute and chronic conditions.

But these were arguments based on implausibility and appeal to Occam's razor. In Vermont, as discussed above, Wennberg and Fowler were able to obtain direct evidence for the lack of a role of illness and patient behavior in seeking care in explaining Vermont's geographic variations. The 1998 Dartmouth Atlas of Health Care (Wennberg and Cooper *1998*) provided several tests of the role of illness in determining utilization, all with "negative" results. Regional variation in surgical procedures for cardiovascular disease across the USA was shown to have little relationship to underlying incidence of cardiovascular disease, as measured by incidence of heart attacks and strokes. While sicker patients (measured by self-reported illness) did indeed use more care than those who were less sick, the rates of use of health care were greater for those with the same reported health states in regions with a high supply of beds compared to those with a low supply, as can be seen in Fig. 8.

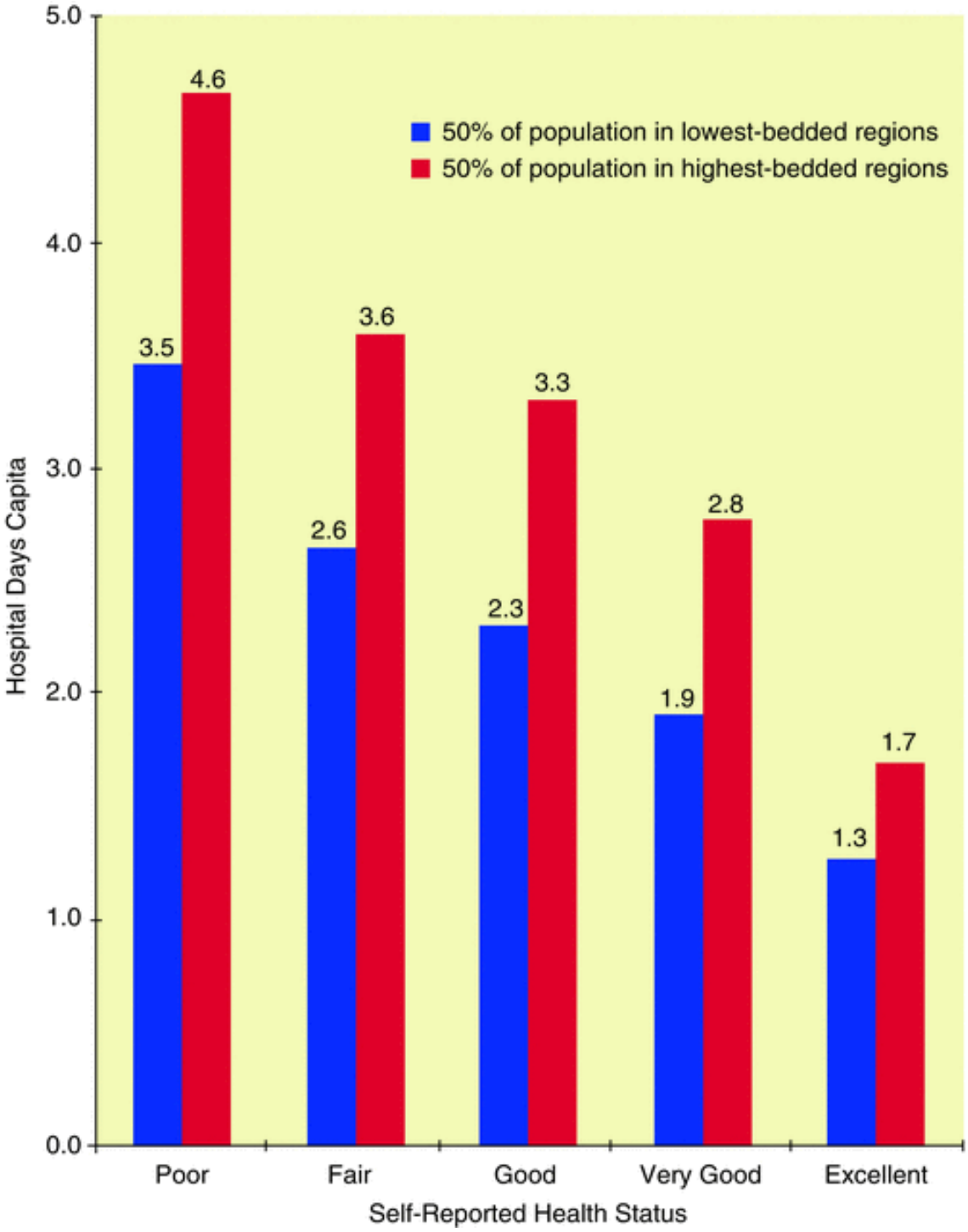


Fig. 8

Self-reported health status and hospital days segmented by regions with high and low supplies of hospital beds (1993). The left-hand (*blue*) bars represent the population living in the hospital referral regions with low per capita supplies of hospital beds; the right-hand (*red*) bars represent those in hospital referral regions with high per capita supplies of hospital beds. The vertical axis is the average number of days spent in hospitals; the horizontal axis is self-reported health status. Medicare enrollees living in regions with higher per capita supplies of hospital beds had higher hospital use, independent of reported health status (Source: the Dartmouth Atlas of Health Care *1998*)

In yet another approach to account for illness, the Dartmouth team conducted a statistical study to progressively adjust the crude regional Medicare spending rates in the 306 Dartmouth Atlas referral regions, first for age, sex, and race and then for age, sex, race, and illness. Adjustment for age, sex, and race had very little effect on the variation, reducing variation to only 4 %. Five of the low-variation conditions/procedures that small-area analysis suggests are driven primarily by illness and were used to adjust the rates: hip fracture, surgery for cancer of the colon or lung, gastrointestinal bleeding, myocardial infarct, and stroke. Further adjustment of age-, sex-, and race-adjusted spending for these illness measures resulted in only a 13 % total reduction in variation in unadjusted spending. Not much of the more than threefold variation appeared to be explained by this aggregated illness measure (Wennberg and Cooper *1998*).

The role of illness, poverty, and other socioeconomic factors in determining demand continues to be a hot topic of debate, and the Dartmouth faculty continues to investigate robust methods of risk adjustment. More recent research has uncovered important flaws in the risk adjustment methods that depend on diagnoses reported in claims data to measure illness (Song et al. *2010*; Welch et al. *2011*; Wennberg et al. *2013*, *2014*). These measures are employed by many outcomes researchers and by the US Medicare program to adjust payments or quality indicators for illness. The problem is that the frequency of diagnosis is not related just to illness; it is also related to the capacity and utilization rates of the health-care system. Patients in regions with more physicians receive more physician visits and referrals; more encounters lead to more diagnoses. As a result, when these methods are used for adjustment, patient populations in high-capacity/high-intensity care regions, such as Miami, appear sicker than they would have had the intensity of care been less, while those in low-capacity/low-intensity regions, such as Seattle, appear correspondingly healthier. Because payment is based in part on this measured "sickness" level, the process results in higher payments to providers in Miami and other regions with high-intensity pattern of care. This might explain at least part of the "inverse care law" – more care is provided to populations with lower needs.

Arguably, the most successful Dartmouth strategy for controlling for illness is the follow-back from death model adopted for evaluating the intensity of use of supply-sensitive care. In this strategy, the Medicare database is used to accumulate all claim records of use of service over fixed intervals prior to date of death and then to compare the care received according to the region of decedent residence or according to the hospital the decedent most often used. Measures of supply-sensitive care include physician visits, referrals, hospitalizations, stays in ICUs, medical imaging studies, and laboratory tests. This strategy has proved very useful in evaluating the intensity of use of supply-sensitive care over given periods of time in the progression of chronic illness and evaluating end-of-life care (Wennberg et al. *2005*; Wennberg *2010*).

Capacity and Supply-Sensitive Care

The New England studies, particularly those of Boston-New Haven, left a strong impression that the size of a hospital relative to the size of the population it served was an important driver of variation in hospitalization rates for acute and chronic medical conditions. Based on conversations with clinicians, the influence of capacity on clinical decision-making appeared to be subliminal. It wasn't perceived as something that affected their decision to admit patients to hospital, even among those academic physicians who had practiced in both communities. The 1998 Dartmouth Atlas revealed that the New England story also held for the nation as a whole. The reports uncovered striking variation in the supply of hospital beds and a strong relationship between supply and hospitalization rates for medical conditions (see Fig. 9). There was also a strong relationship between the supply of medical specialists and the rate of use of their services.

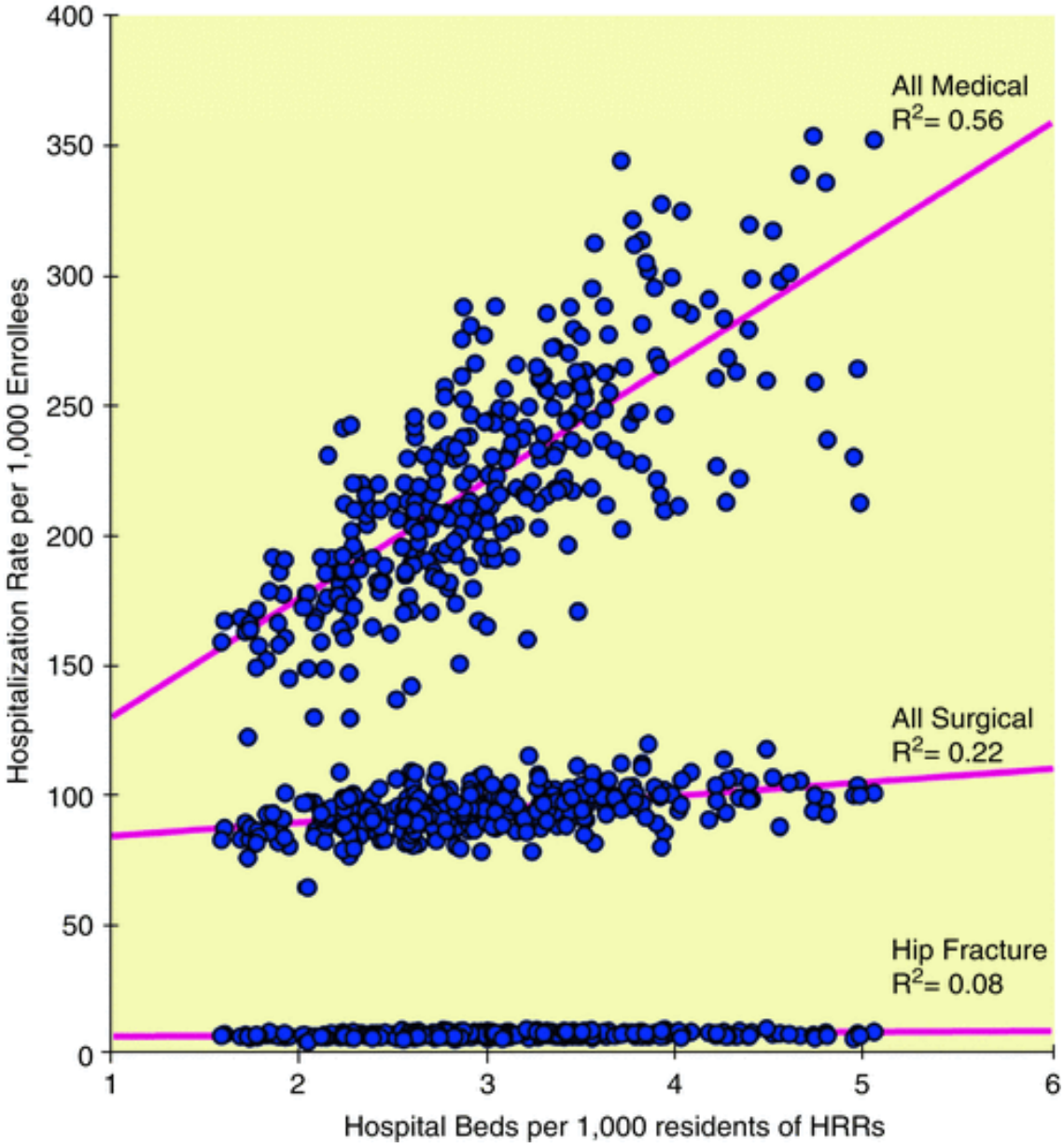


Fig. 9

The association between allocated hospital beds and Medicare hospitalizations for medical hospitalizations, surgery, and hip fracture. The hospitalization rate for medical conditions is strongly correlated with bed supply ($R^2 = 0.56$), surgical hospitalizations are less strongly correlated ($R^2 = 0.22$), and hip fracture hospitalizations have little correlation ($R^2 = 0.08$) (Source: the Dartmouth Atlas of Health Care *1998*)

A simple description may help in understanding how capacity influences utilization. For historic/political reasons, which seem to have little to do with illness or socioeconomic characteristics, the number of hospital beds and physicians on a per capita basis is variably distributed across hospital service areas. Medical science, as can be confirmed by analysis of practice guidelines, medical textbooks, and the research literature, is relatively silent concerning when to hospitalize or admit to an ICU, when to schedule a revisit or refer to a specialist, or when to order an MRI or a laboratory test. Thus, medical science does not affect relative supply importantly. In the absence of information from medical science, the widely held cultural assumption that more care is better leads naturally to

physicians using available resources up to their limits, albeit subconsciously. Under such circumstances, variations that relate to capacity are to be expected.

This is a different model for the role of supply in inducing demand than that posited by clinical agency theory. Instead of the marginal influence of deviant physicians who induce demand for self-serving reasons, it highlights the well-meaning tendency to use whatever resources are available, but to do so without evidence of marginal benefit, or even awareness of the effects of local supply, even at the most prestigious academic medical centers. (It is also a different model than that posited for preference-sensitive treatments – see below.)

The economic consequences of supply-sensitive care are significant. It is the frequency of use of this category of care that accounts for the more than twofold regional variation in Medicare spending across the USA (*Wennberg et al.*). Most of these resources are used in managing those with chronic illness, with over 30 % of total spending allocated to those who are in the last 2 years of life. The consequences for the patient experience are also significant. Capacity dominates and determines the intensity of care, even when, as is the case for end-of-life care, patient preferences should play a role. The variation in intensity of care at the end of life is illustrated dramatically by the five academic medical centers that US News and World Report claims to be America's best (see Fig. 10). A reformed science of health-care delivery would surely ensure that academic medical centers take responsibility for rationalizing the striking variation in their own practices and build a scientific basis for managing chronic illness. This should be viewed as an essential priority for addressing unwarranted variation.

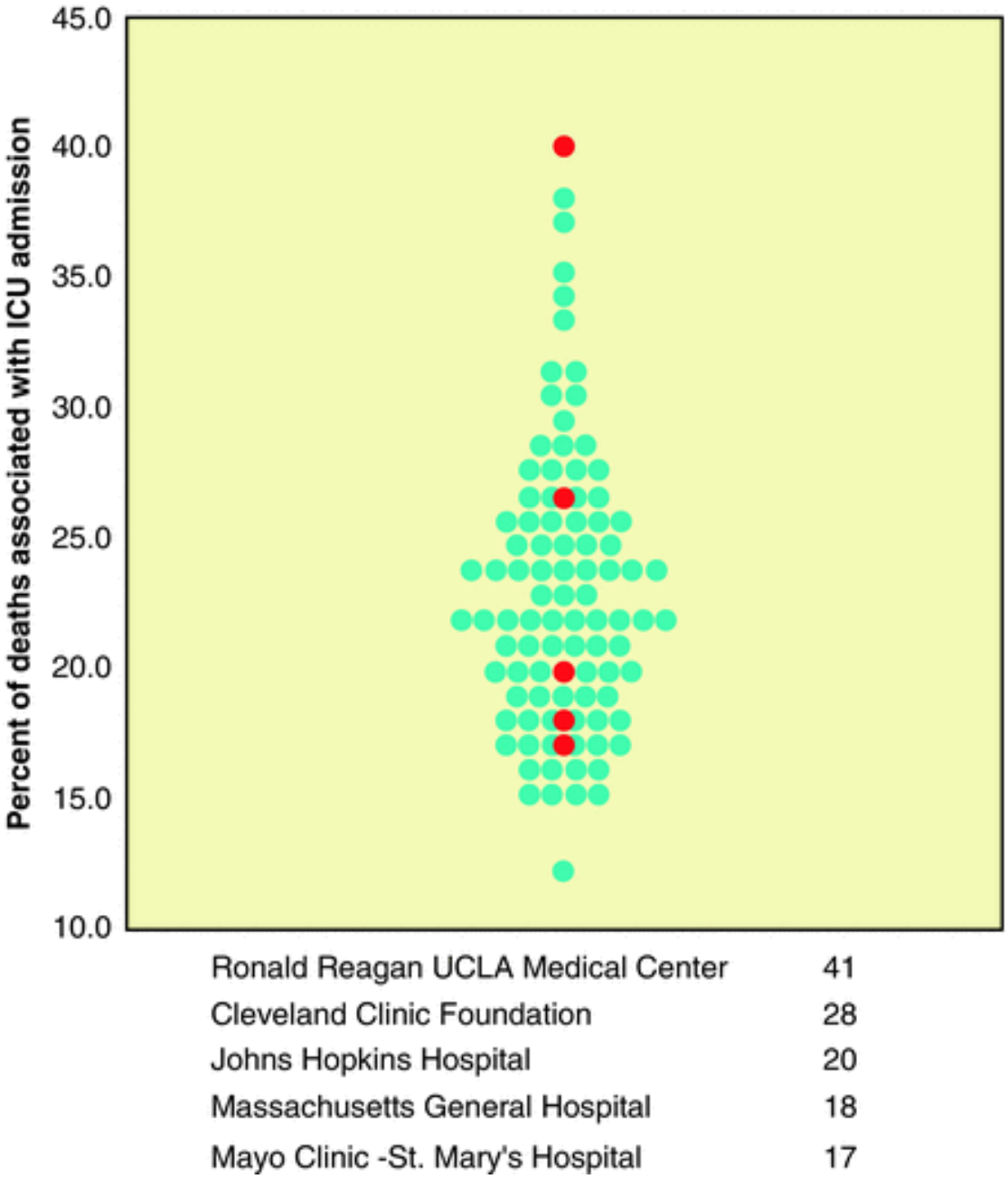


Fig. 10

Percent of patients with chronic illness who experienced a stay in an ICU at the time of death. The data are for patients who received most of their care during the last 2 years of life from an academic medical center associated with a US medical school. Data are for deaths in 2010. *Red dots* indicate the five best US hospitals according to *US News* for 2011–2012 (Source: unpublished data from the Dartmouth Atlas Project for deaths that occurred in 2010)

However, reform of science policy alone will not be enough to stem the dynamics of growth and stabilize the health-care economy. Here, the Dartmouth work leads to the necessity to directly influence capacity. Greater intensity of supply-sensitive care doesn't improve life expectancy (Fisher et al. *2003*), and the negative consequences of maintaining the current irrational regional distribution of medical capacity appear to outweigh the benefits. The Dartmouth Atlas has uncovered a number of relatively efficient health-care organizations that provide benchmarks for resource use upon which

strategies to limit capacity (such as through health planning) or spending (such as prospective budgeting or capitation) can be based. They also provide benchmarks for the design of a physician workforce tailored to the needs of such organizations.

Medical Opinion and Preference-Sensitive Care

A few years after Wennberg and Gittelsohn published their article in *Science*, the Wennberg team undertook a review of the scientific literature to document clinical controversies for common surgical procedures and to see if there was a link between the degree of controversy surrounding a procedure and the degree to which its use varied among small areas (as illustrated in Fig. 2). The study uncovered extensive professional disagreement on reasons why particular operations should be performed – on such basic issues as whether the purpose of the procedures was to improve the quality of life or to extend life expectancy (Wennberg et al. *1980*). They found evidence that the more extensive and "hot" the controversies, the greater the degree of geographic variation. They also found extensive disagreement as to what the outcomes actually were and their chances of happening. The implications were summarized as follows:

The limits of informed decision-making in medical markets are more severe than is generally realized... If the outcomes of alternative treatments are not understood, how is it possible to make informed decisions or give informed consent? ... When decisions are made, whose values are being expressed, the patient's or the physician's? The geographic variations in exposure rates [to a procedure] are consistent with the thesis that medical care choices are highly dependent on the preferences of physicians. When professional disagreement is strong and patients delegate decision-making to physicians, the probability of exposure to specific interventions will often depend on the style of practice or clinic selected for care rather than the nature and severity of illness.

The opportunity for the Dartmouth team to directly investigate the relationship between scientific uncertainty over surgical theory and practice variation – and to illustrate the importance of linking variations research to outcomes research – arose in the late 1970s when information on surgical variation among Maine hospital service areas was made available to physicians throughout the state. The information had a far-reaching impact, thanks in large part to the leadership of Dr. Daniel Hanley, the editor of the Journal of the Maine Medical Association, which had published several articles by Wennberg and Gittelsohn. He organized practicing physicians to work together with the research team to address the clinical reasons why surgery varied so much. The discussions with Maine's urologists as to why surgery for an enlarged prostate varied sixfold among hospital service areas led to a 20-year outcomes research project that resulted in a revision in the underlying clinical theory for undertaking surgery for this condition. The urologists disagreed among themselves as to why the procedure should be undertaken. Some ascribed to the preventive theory of surgery for benign prostatic hypertrophy or "BPH" – operate early to prevent progression of the disease, which threatens to block the bladder and kidney and reduces life expectancy. Others believed that the course of untreated BPH was not so virulent; for most patients, the reason they did surgery was to reduce bothersome symptoms and improve the quality of life.

Through a series of studies, summarized in Tracking Medicine (Wennberg *2010*) and in annotated form on the Dartmouth Institute for Health Policy and Clinical Practice website (*Wennberg Anthology*), the research team established that the preventive theory didn't hold up. Long-term survival wasn't improved by surgery. The procedure did result in a significant improvement in urinary symptoms, and for some men this meant a significant improvement in the quality of life. However, the operation was also associated with changes in sexual performance that reduced the quality of life for

some men. Thus, the "right treatment" for the individual patient depended on how he weighed the risks and benefits; in other words, it depended on his preferences. This, in turn, depended on reform of the doctor-patient relationship from the delegated decision-making model supported by clinical agency theory to a shared decision-making model designed to promote informed patient choice.

As the research project matured, the research team began to experiment with the use of decision aids to help patients and their physicians participate in shared decision-making. Evidence emerged that informed patients made better decisions. In the case of BPH, they understood what was at stake – the gist of the decision problem they faced. Informed patients also chose their treatments more wisely. Those who were concerned about preserving sexual function were much more likely to avoid surgery, while those concerned most about relief from symptoms of BPH chose surgery. It was also learned that shared decision-making can change the demand for surgery, at least when implemented in a capitated or budgeted health-care system. Colleagues at Group Health in Seattle and Kaiser-Denver undertook an experiment to implement shared decision-making for BPH surgery throughout their systems, resulting in a dramatic drop in the population-based rates of surgery.

Promoting an active role for the patient in the choice of treatment should become a central strategy to reform health care. Outcomes research and evidence-based medicine are not enough. It is clear that medical opinion exercises its influence on utilization through two channels. One concerns the assumptions about the theory for doing a treatment and what the expected outcomes of care are – in short, opinion about what works. The other concerns opinion about the value of a specific treatment to the individual patient – opinion about what patients want. Thus, reducing unwarranted influence of medical opinion on utilization for preference-sensitive conditions requires not only research to reduce scientific uncertainty about theory and outcomes; it also requires establishing a doctor-patient relationship based on shared decision-making, one that ensures that the values of the patient rather than the opinion of the physician determine demand (utilization). In short, available treatment options should be evidence based, but medical necessity should be determined by informed patient choice.

The importance of informed patient choice in reducing unwarranted variations in preference-sensitive care has broad implications for future research in the heath-care delivery sciences. These include questions concerning the measurement and stability of patient preferences; the design and evaluation of decision aids designed to promote informed patient choice, including the study of framing effects, methods of informing patients about risk, and the measurement of decision quality; and the conduct of clinical outcomes research following cohorts of patients where choice of treatment is based on informed patient choice.

Conclusions

A central objective of public policy for health is to facilitate the provision of appropriate care – especially in response to patient preferences – in a health-care "market" characterized by a very unequal distribution of information between physicians and patients. But work on variations indicates that even the experts are not as knowledgeable as might be required to properly fulfill this objective. In the case of preference-sensitive interventions such as elective surgery or diagnostic screening exams, variation has been traced to differences in medical opinion that arise most fundamentally because of inadequate knowledge of the outcomes of treatment options and flaws in clinical agency that result in a misdiagnosis of the individual patient's preferences. The general outline of a strategy for reducing unwarranted variation would link the monitoring of variation to a well-funded program in outcomes and preference research. It would also seek to promote shared decision-making/informed

patient choice to replace delegated decision-making/clinical agency as the ethical standard of practice for preference-sensitive care.

The ubiquitous patterns of variation in common preference-sensitive surgical procedures have been examined, with the conclusion that they mostly represent a massively inefficient use of health-care resources. But the general point is that the identification of that inefficiency is often unknown, in the sense that the right rate is unclear because the knowledge base on which to make proper judgments, taking account of patient preferences, is immature. As shared decision-making replaces delegated decision-making, and outcomes research increases knowledge, more benchmarks can be anticipated for the demand for preference-sensitive care that is driven by informed patient choice.

Proper analyses of research priorities identified by the implications of this lack of knowledge are essential and should play an important role in the appropriate development of guidelines.

Unwarranted variation requires serious study in different contexts and analysis of their implications on health and costs and the possibility of research providing enlightenment. For example, the evidence recounted above suggesting a convergence of unwarranted variation for hysterectomy and appendectomy between OECD countries requires further study.

In the case of supply-sensitive care, where the issue is the frequency or intensity of everyday care in managing illness, a strategy for reducing unwarranted variation would include the control of capacity through health planning of resources or regulation of budgets. At the same time, health policy would promote organized systems of care capable of managing care of patient populations over time, learning from that experience to improve care and help patients achieve their goals, particularly for end-of-life care.

In contrast to preference-sensitive care for which no population rate is right, the analysis of variation in supply-sensitive care supports the identification of inefficiency, at least in the US context. The Dartmouth Atlas uncovered a number of low-cost high-quality health-care organizations that provide benchmarks for the efficient use of supply-sensitive care. These benchmarks, which predict savings of some 30 % of Medicare spending in managing patients with chronic illness over the course of their illness, could be used in strategies to limit capacity through health planning or spending targets. These organizations also provide benchmarks for an efficient physician workforce.

Variation in effective care remains the simplest construct of unwarranted variation, but one that continues to challenge efforts to remedy widespread shortcomings in medical practice. The movement to improve technical quality where benefits clearly outweigh potential harm has led to the growth of diverse efforts worldwide to improve this type of care. Progress has been made on some fronts, but implementation of clearly beneficial care and avoidance of harmful care remains incomplete (McPherson and Bunker *2007*).

A strategy for reducing unwarranted variation will require the continuous availability of population-based information describing health-care delivery. In the absence of such activity, there is no knowledge of the extent of variation among neighboring medical communities or administrative areas, even among frontline clinicians and administrators, much less patients. Feedback of such data is essential to get the attention of patients, practicing physicians, policy makers, medical leaders, politicians, the press, and the general public. Under favorable circumstances, this can lead to constructive engagement in strategies to reduce unwarranted variation.

The recent trend to publish atlases of health care (patterned after the Dartmouth Atlas of Health Care) in Spain, the Netherlands, Canada, Australia, and England is an important step in creating transparency that should be continued and expanded. Their role in health planning and commissioning, where evaluated properly, is vital and an essential public policy intervention, given the flaws in the clinical agency model. But these reports and their analyses need to be based on peer-reviewed research. The measures of variation and other aspects of methodology employed in atlases of health care should be rigorous and consistent with OECD methods currently being developed. It is

important to avoid idiosyncratic measures of variation that do not conform to epidemiologic standards (Department of Health *2010*). Measures of systematic variation, adjusted where possible for demographic differences and properly adjusted for morbidity, are essential, not measures that are falsely inflated by the supply of resources.

The OECD 13-country study stands as a landmark in international cooperation and resolves in addressing the problem of unwarranted variation in health-care delivery. The analysis of the causes and consequences of variation as an international collaboration should continue. National research funding bodies should be fully involved in developing a coherent research agenda. There is a large area of as yet insufficient research on the role of informed patient preference and the proper assessment of outcomes attributable to varying intervention rates. Study of the comparative cost-effectiveness of high versus low rates of delivery of supply-sensitive care should also be given emphasis. Within most OECD countries, hospitalizations for medical conditions vary twofold among regions, about as much variation as seen within the USA. How many of these regions follow the low-cost high-quality profile and thus qualify as a benchmark for efficiency? And how much savings do they predict would follow their adoption as the standard of practice? These as yet unanswered questions would seem to be important to patients, health-care professionals, and policy makers in all countries.

Electronic supplementary material

Video 1

WINDOWS ON HEALTHCARE

MOV file: 3030384 kb

References

Appleby J, Raleigh V, Frosini F, et al. Variations in health care. The good, the bad and the inexplicable. London: King's Fund; 2011.

Bunker JP. Surgical manpower: a comparison of operations and surgeons in the United States and in England and Wales. N Engl J Med. 1970;282:135–44.
CrossRef PubMed

Burton M. Tonsillectomy – then and now. Int J Epidemiol. 2008;37:23–5.
CrossRef PubMed

Department of Health. The NHS atlas of variation in healthcare. 2010. http://www.rightcare.nhs.uk

Dyck FJ, Murphy FA, Murphy JK, et al. Effect of surveillance on the number of hysterectomies in the province of Saskatchewan. N Engl J Med. 1977;2:1326–8.
CrossRef

Fisher ES, Wennberg DE, Stukel TA, et al. The implications of regional variations in Medicare spending: part 2. Health outcomes and satisfaction with care. Ann Intern Med. 2003;138:288–98.
CrossRef PubMed

Glover JA. The incidence of tonsillectomy in school children. Proc R Soc Med. 1938;31:1219–36. Reprinted in the Int J Epidemiol. 2008;37:9–19.
PubMedCentral PubMed

Kuhn TS. The structure of scientific revolutions. 50th anniversary. Ian Hacking (intro.). 4th ed. University of Chicago Press; 2012. p. 264. ISBN 978-0-226-45811-3. LCCN 2011042476.

Lembcke PA. Measuring the quality of medical care through vital statistics based on hospital service areas: 1. comparative study of appendectomy rates. Am J Public Health Nations Health. 1952;42(3):276–86. doi:10.2105/AJPH.42.3.276.
PubMedCentral CrossRef PubMed

Lewis CE. Variations in the incidence of surgery. N Engl J Med. 1969;281:880–4.
CrossRef PubMed

Lichner EA, Pflanz M. Appendectomy in the Federal Republic of Germany: epidemiology and medical care patterns. Med Care. 1971;9:311–8.
CrossRef

McPherson K. International differences in medical care practices. In: International comparison of health care financing and delivery: data and perspectives. Health Care Financing Administration; Washington: U.S. Government Printing Office, Dec. 1989, (Annual Supplement) pp. 9–20.

McPherson K, Bunker JP. Costs, risks and benefits of surgery: a milestone in the development of health services research. J R Soc Med. 2007;100:387–90.
PubMedCentral CrossRef PubMed

McPherson K, Wennberg JE, Hovind OB, Clifford P. Small-area variations in the use of common surgical procedures: an international comparison of New England, England and Norway. N Engl J Med. 1982;307:1310–4.
CrossRef PubMed

McPherson K, Downing A, Buirski D. Systematic variation in surgical procedures and hospital admission rates, PHP publication no 23. London: London School of Hygiene and Tropical Medicine; 1996.

McPherson K, Gon G, Scott M. International variations in a selected number of surgical procedures. OECD Health Working Papers, No. 61, OECD Publishing; 2013. doi:10.1787/5k49h4p5g9mw-en.

OECD Report. Geographic variations in health care: what do we know and what can be done to improve health system performance? OECD Health Policy Studies, OECD Publishing; 2014. doi:10.1787/9789264216594-en.

Scott M, McPherson K, McKenzie I, et al. International variations in rates of a range of surgical procedures. Report to the Foundation for Informed Medical Decision Making. University of Oxford; 2008.

Song Y, Skinner J, Bynum J, et al. Regional variations in diagnostic practices. N Engl J Med. 2010;363:45–53.
PubMedCentral CrossRef PubMed

Vayda E. A comparison of surgical rates in Canada and in England and Wales. N Engl J Med. 1973;289:1224–9.
CrossRef PubMed

Welch HG, Sharp SM, Gottlieb DJ, et al. Geographic variation in diagnosis frequency and risk of death among Medicare beneficiaries. JAMA. 2011;305(11):1113–8.
PubMedCentral CrossRef PubMed

Wennberg JE. Tracking medicine: a researcher's quest to understand health care. New York: Oxford University Press; 2010. Chapters 6 and 7.

Wennberg Anthology. 2015. http://www.wennberganthology.org/the-prostate-papers-outcomes-research-and-the-pathway-to-shared-decision-making/.

Wennberg Anthology. 2015. http://www.wennberganthology.org.

Wennberg JE, Cooper MM, editors. The Dartmouth Atlas of Health Care 1998. Chicago: American Hospital Publishing; 1998.

Wennberg J, Fowler FJ. A test of consumer contributions to small area variations in health care delivery. J Maine Med Assoc. 1977;68:275–9.
PubMed

Wennberg J, Gittelsohn A. Small area variations in health care delivery: a population-based health information system can guide planning and regulatory decision-making. Science. 1973;182:1102–8.
CrossRef PubMed

Wennberg J, Gittelsohn A. Health care delivery in Maine I: patterns of use of common surgical procedures. J Maine Med Assoc. 1975;66:123–30, 149.
PubMed

Wennberg J, Gittelsohn A. Variations in medical care among small areas. Sci Am. 1982;246(4):120–34.
CrossRef PubMed

Wennberg JE, Fisher ES, Skinner JS. Geography and the debate over Medicare reform. 2002. Web-Exclusive on Health Affairs. http://content.healthaffairs.org/cgi/content/full/hlthaff.w2.96v1/DC1

Wennberg J, Blowers L, Parker R, Gittelsohn AM. Changes in tonsillectomy rates associated with feedback and review. Pediatrics. 1977;59:821–6.
PubMed

Wennberg JE, Bunker J, Barnes B. The need for assessing the outcomes of common medical practices. Annu Rev Public Health. 1980;1:277–95.
CrossRef PubMed

Wennberg JE, McPherson K, Caper P. Will payment based upon diagnosis-related groups control hospital costs? N Engl J Med. 1984;311:295–300.
CrossRef PubMed

Wennberg JE, Freeman JL, Culp WJ. Are hospital services rationed in New Haven or over-utilized in Boston? Lancet. 1987;1(8543):1185–8.
CrossRef PubMed

Wennberg JE, Fisher ES, Baker L, Sharp SM, Bronner KK. Evaluating the efficiency of California providers in caring for patients with chronic illness. Web-Exclusive on Health Affairs. 2005. http://content.healthaffairs.org/cgi/content/full/hlthaff.w5.526/DC1

Wennberg JE, Staiger DO, Sharp SM, et al. Observational intensity bias associated with illness adjustment: cross sectional analysis of insurance claims. BMJ. 2013;346:f549.
PubMedCentral CrossRef PubMed

Wennberg DE, Sharp SM, Bevan G, et al. A population health approach to reducing observational intensity bias in health risk adjustment: cross sectional analysis of insurance claims. BMJ. 2014;348:g2392.
PubMedCentral CrossRef PubMed