Comparing Hospital Processes and Outcomes in California Medicare Beneficiaries: Simulation Prompts Reconsideration

Gabriel J Escobar, MD; Jennifer M Baker, MPH, CHES; Benjamin J Turk, MAS; David Draper, PhD; Vincent Liu, MD, MS; Patricia Kipnis, PhD

Perm J 2017;21:16-084

E-pub: 10/05/2017

https://doi.org/10.7812/TPP/16-084

ABSTRACT

Introduction: This article is not a traditional research report. It describes how conducting a specific set of benchmarking analyses led us to broader reflections on hospital benchmarking. We reexamined an issue that has received far less attention from researchers than in the past: How variations in the hospital admission threshold might affect hospital rankings. Considering this threshold made us reconsider what benchmarking is and what future benchmarking studies might be like. Although we recognize that some of our assertions are speculative, they are based on our reading of the literature and previous and ongoing data analyses being conducted in our research unit. We describe the benchmarking analyses that led to these reflections.

Objectives: The Centers for Medicare and Medicaid Services' Hospital Compare Web site includes data on fee-for-service Medicare beneficiaries but does not control for severity of illness, which requires physiologic data now available in most electronic medical records.

To address this limitation, we compared hospital processes and outcomes among Kaiser Permanente Northern California's (KPNC) Medicare Advantage beneficiaries and non-KPNC California Medicare beneficiaries between 2009 and 2010.

Methods: We assigned a simulated severity of illness measure to each record and explored the effect of having the additional information on outcomes.

Results: We found that if the admission severity of illness in non-KPNC hospitals increased, KPNC hospitals' mortality performance would appear worse; conversely, if admission severity at non-KPNC hospitals' decreased, KPNC hospitals' performance would appear better.

Conclusion: Future hospital benchmarking should consider the impact of variation in admission thresholds.

INTRODUCTION

When people book a commercial airline flight, they expect more than a safe ride. They can compare prices and find information on flight delays or lost luggage. Further, they can expect that, in general, the service they receive from a given airline will be similar whether they buy their ticket in Denver, CO, or Lexington, KY. Contrast this with the current situation in health care, where information on what actually happens to patients—whether in terms of safety, quality, cost, or service—that would permit a consumer to make an informed choice is much more scarce, and where processes and outcomes for a given procedure or illness vary dramatically from site to site, even within a single system. This lack of transparency even affects clinicians but is particularly true for consumers without formal medical training.

One major difference between these two industries, of course, is that the end products are different. Generally speaking, once a customer has paid for an airline ticket, it is unlikely that his or her past travel history will have a huge impact on the flight time between Denver, CO, and Chicago, IL, or on the airline's revenue. On the other hand, this is clearly not the case with respect to a patient's medical history when someone boards an ambulance or walks into a clinic. In the assessment of health care quality, risk adjustment—accounting for baseline risk—is critical.

During the last few decades, a slow but continued transformation has been occurring in medicine, one that seeks to move institutions to save lives and decrease suffering through the improved use of information. This transformation has led to the development and publication of risk-adjusted outcome reports that compare and benchmark the institutions' performance. Risk-adjusted benchmarking is primarily a branch of health services research, itself a descendant of epidemiology. It has a much lower public profile than do other branches of medicine. Although risk-adjusted benchmarking is largely driven by statistics and informatics, its practitioners must also consider organizational psychology and political science to carry out their work.

Risk-adjusting techniques are becoming increasingly important as more medical care in the US takes place in integrated systems. Quality-assurance departments can examine and compare what processes are in place at high-ranking centers that differ from processes at lower-ranking centers. When these comparisons are conducted properly and are supported by political will, discoveries and improvements follow. This approach to process improvement has a strong track record in multiple areas of medicine, ranging from the care of newborns[1] to that of adults.[2-4] Using rigorous risk adjustment methods for ranking institutions is crucial to ensure that interinstitutional differences are attributed to differences in processes and not differences in the characteristics of the

Gabriel J Escobar, MD, is the Regional Director for Hospital Operations Research for The Permanente Medical Group, Inc, at the Division of Research in Oakland, CA. E-mail: gabriel.escobar@kp.org. **Jennifer M Baker, MPH, CHES,** is a Public Health Program Specialist for Contra Costa Public Health Clinic Services in Martinez, CA. E-mail: calhounjennifer1@gmail.com. **Benjamin J Turk, MAS,** is a Data Analyst for the Division of Research in Oakland, CA. E-mail: benjamin.j.turk@kp.org. **David Draper, PhD,** is a Professor of Applied Mathematics and Statistics at the University of California, Santa Cruz. E-mail: draper@soe.ucsc.edu. **Vincent Liu, MD, MS,** is the Regional Director for Hospital Advanced Analytics for The Permanente Medical Group, Inc, at the Division of Research in Oakland, CA. E-mail: vincent.x.liu@kp.org. **Patricia Kipnis, PhD,** is the Principal Statistician for Decision Support at Kaiser Foundation Health Plan. E-mail: patricia.kipnis@kp.org.

underlying population. Thus, in the best scenario, benchmarking can save lives and decrease morbidity, which is a major reason health services researchers devote so much effort to improving risk adjustment methods. However, this does not mean that benchmarking processes are uniformly useful (or similar), and substantial concerns exist regarding both the proliferation and the quality of individual benchmarking systems.[5,6] Those working in this field recognize that one *should* entertain doubts about what we do, particularly given both practical and theoretical concerns about the limitations of the methods we employ.[7,8]

One notable example of this expansion is a recent set of hospital rankings (*US News Best Hospitals*) issued by *US News & World Report*.[9,10] The new rankings are remarkable for two reasons: One that would be easily comprehensible to most of its readers, and one that is less obvious. The "easy" reason is that, unlike most rankings in the popular press (and unlike its previous incarnations in the magazine), the more recent set of rankings use objective, publicly available data rather than "expert" opinion. (Rankings based on this approach have poor correlation with patient outcomes.[11]) The less obvious reason is that, unlike benchmarking done by (to give one important example) the federal government's Centers for Medicare and Medicaid Services' (CMS) Hospital Compare Web site,[12] this particular report included not just data on fee-for-service Medicare beneficiaries. Remarkably, they now also include data from some large managed care providers, which are not available from the Medicare data warehouse. Non-fee-for-service patients constitute approximately 31% of all beneficiaries[13] and are referred to as Medicare Advantage patients.

In this article, we reflect on a benchmarking study described in the Sidebar: Benchmarking Study on Effects of Including Physiologic Severity Adjustment. The findings of this study led us in a different direction than originally anticipated. We started with one question: Compared with the outcomes of other patients in California, once a patient enters one of Kaiser Permanente Northern California's (KPNC's) hospitals, how does s/he fare? We ended up asking two other, more speculative, questions. What factors affect the decision to admit a patient to the hospital in the first place? How would one study these factors?

STUDY SETTING: HOW WE GOT TO OUR INITIAL QUESTION

The setting for our work is KPNC, a capitated integrated health care delivery system. Under a mutual exclusivity agreement, 9500 salaried physicians of The Permanente Medical Group, Inc, provide care for 4.1 million members of Kaiser Foundation Health Plan, Inc, at more than 200 clinical care locations, which include 21 hospitals operated by Kaiser Foundation Hospitals, Inc. Its comprehensive information systems—built around a common medical record number—permit KPNC to track information throughout the continuum of care, including care covered by the Health Plan but delivered elsewhere.[14] The Epic (Epic Systems Corp, Intergalactic, Verona, WI, www.epic.com) electronic medical record (EMR), known internally as KP HealthConnect, was fully implemented during a 5-year period ending in 2010.

In our department, the KPNC Division of Research, our team focuses on the outcomes of hospitalized adults. It is known that KPNC gets high marks for quality in multiple areas.[15,16] However, outside Kaiser Permanente, much less is known about the intense degree of self-examination conducted by internal KPNC departments. In our specific area of expertise, which includes risk adjustment of hospital outcomes,[17-22] our work has focused on addressing variation in mortality across hospitals.

For a layperson, the term *practice variation,* when not defined in research terms, might refer to something inevitable and innocuous. It makes intuitive sense that a patient who was in a major motor vehicle crash who is treated at a small hospital might do worse than one treated at a major trauma center; similarly, common sense would suggest that a patient with pneumonia in both lungs and a bloodstream infection (sepsis) will do worse than a patient with "walking" (mild) pneumonia. It also makes intuitive sense that individual physicians' practicing "styles" differ; such variation may in fact be desirable. However, when health services researchers talk about practice variation, they are talking about a far more problematic and insidious issue—the fact that processes, costs, and outcomes for *very similar* patients vary across similar institutions. Moreover, this variation, sometimes referred to as *residual* variation, persists after statistical adjustment for many patient characteristics. Consequently, a major proportion of the efforts of both health services researchers and health care quality assurance teams focuses on identification of best practice and, optimally, eliminating variation from best practice.

RECONSIDERATION OF BENCHMARKING

There are limitations to the analyses described in the Sidebar: Benchmarking Study on Effects of Including Physiologic Severity Adjustment. Our restriction to the ten conditions accounting for most of the deaths means that our analyses may not apply to conditions that have relatively low mortality but which may have high morbidity and/or cost. Admissions data used to estimate the likely severity of illness are based only on KPNC data. Given lack of such data from non-KPNC hospitals, this was a reasonable step, but one can question the approach and the degree of generalizability.

Going into this study, we knew that measurement of physiologic severity of illness would have an effect because it is known that adding clinical data to hospital risk adjustment has a huge impact.[23,24] Multiple studies, including our own, suggest that at least one-half of the predictive ability of models that predict hospital mortality comes from physiologic measures,[17,25] which is a major reason severity scores have face validity among clinicians. Presumably, if hospitals functioned as isolated entities, the distribution of their patients' severity of illness at admission would be a direct reflection of the general health of their local area. However, and this is particularly true in California, where many hospitals now function as parts of systems with varying degrees of integration, the severity of illness distribution is now likely to be shaped by two factors. The first is the degree to which systems are in place to prevent hospitalization. An increasing proportion of processes that formerly required hospitalization can now be handled on an outpatient basis, as can be seen, for example, with congestive heart failure.[26] In general, the more effective a health care organization is in deploying such systems, the more likely it

BENCHMARKING STUDY ON EFFECTS OF INCLUDING PHYSIOLOGIC SEVERITY ADJUSTMENT

INTRODUCTION

The Centers for Medicare and Medicaid Services (CMS) Hospital Compare Web site (www.medicare.gov/hospitalcompare/) provides risk-adjusted comparisons of processes and outcomes of hospitalized Medicare patients using a transparent and reproducible method developed by the team of Harlan Krumholz, MD, at Yale University in New Haven, CT.[1,2] Analyses based on the basic method have yielded insights on implementation of quality improvement projects.[3,4]

Existing CMS data transfer protocols are such that a substantial proportion of Medicare members, Medicare Advantage patients, are not included in the analyses; only fee-for-service Medicare members are included. Almost all (99.8%) Kaiser Permanente Northern California (KPNC) Medicare members are in the Medicare Advantage program. Consequently, the sample for KPNC hospitals in the Hospital Compare Web site is extremely small and nonrepresentative, and the risk adjustment method sets all KPNC hospitals' performance as "Number of cases too small" or "Not available," effectively eliminating the utility of the Web site for benchmarking KPNC hospital performance. Also, CMS data are limited in that they capture only a patient's acute diagnoses (eg, "this patient is being admitted for acute appendicitis") and comorbid conditions (eg, "this patient also happens to have diabetes and arthritis"). One of the most important dimensions of a patient's illness—admission severity of illness—is not captured. Currently, Hospital Compare is unable to account for this type of severity difference between patients because CMS does not mandate capture of these data. Furthermore, the lag time for CMS data availability (years) exceeds that of KPNC data (1 to 2 months), although 7 of our hospitals are now assigning severity of illness scores in real time.

Comparing our outcomes with those of an external benchmark is highly desirable. Because multiple studies have shown that physiologic data have a large impact on risk adjustment,[5-9] KPNC now routinely employs severity of illness scores using detailed physiologic data (laboratory test results and vital signs). Thus, there is value to estimating what the magnitude of the effect of adjusting physiologic severity might be if these data were included in Hospital Compare.

Table 1. Description of study cohort

Primary condition[a]	CMS[b] Hospitalizations,[d] no.	CMS[b] Mortality, %	KFH Medicare[c] Hospitalizations, no.	KFH Medicare[c] Mortality, %
Hospital Compare AMI	20,076	14.8	2922	12.5
Hospital Compare heart failure	43,424	10.9	5721	12.5
Hospital Compare pneumonia	44,522	12.3	5559	12.8
Sepsis	63,295	28.9	8566	25.2
Acute CVD	28,834	19.7	3877	21.6
Aspiration pneumonia	12,311	27.9	796	37.8
Catastrophic conditions	14,145	27.9	1553	27.0
Other cardiac conditions	34,177	6.7	4226	7.4
Highly malignant cancer	21,501	24.8	2890	24.8
Trauma	54,443	5.4	5539	5.6
Conditions combined	361,728	16.7	42,432	16.6

[a] See Appendix: Part 2 (www.thepermanentejournal.org/files/2017/16-084-Appendix.pdf) for the approach we employed to group individual International Classification of Diseases diagnosis codes into primary conditions.
[b] Refers to California fee-for-service Medicare beneficiaries with claims for the 2009 and 2010 calendar years, excluding all Kaiser Permanente Northern California (KPNC) and Southern California hospitals. Hospitalizations from hospitals with small numbers are included.
[c] Refers to Northern California Kaiser Foundation Health Plan's members with Medicare Advantage coverage for the 2009 and 2010 calendar years. A small number of KPNC fee-for-service members (592 of the combined 35,523 patients) were included in the analyses, as were hospitalizations from hospitals with small numbers.
[d] Refers to the number of *hospitalizations* included in a given analysis; Appendix: Part 1, Table 1.1 (www.thepermanentejournal.org/files/2017/16-084-Appendix.pdf) provides a table in which an individual patient is the unit of analysis.
AMI = acute myocardial infarction; CMS = Centers for Medicare & Medicaid Services; CVD = cardiovascular disease; KFH = Kaiser Foundation Hospitals.

METHODS

We obtained CMS hospitalization data for all non-KPNC California fee-for-service Medicare beneficiaries for the 2009 and 2010 calendar years. Then, with the generous help of Dr Krumholz's team, we formatted KPNC hospitalization data so that the structure was the same, permitting us to merge KPNC data with CMS data. We limited our analyses to the top 10 diagnoses that accounted for 75% of inpatient deaths in 2009 and 2010: Acute myocardial infarction, congestive heart failure, pneumonia, sepsis, stroke, aspiration pneumonia, catastrophic conditions (eg, ruptured aortic aneurysm), other cardiac conditions, malignant cancer, and trauma. Table 1 summarizes the study population serving as the base for our analyses. We replicated the Hospital Compare risk adjustment method (adjusting for patients' age, sex, admission diagnosis, whether hospitalization began in the Emergency Department, and the burden of comorbid illnesses). We then compared processes (use of the Intensive Care Unit, use of assisted ventilation, and length of stay) and outcomes (30-day mortality and 30-day readmission from hospital discharge) across the 323 hospitals in our cohort.

Taking advantage of KPNC's rich information systems, which have permitted us to develop a variety of automated severity measures,[5-8] we assigned hospital records *simulated* severity of illness scores. These scores were imputed on the basis of the reasonable inference that illness severity at admission would tend to be similar in patients with similar characteristics. Using the 2009 and 2010 KPNC hospital cohort, in which each hospitalization record had an admission severity of illness score, called the Laboratory-based Acute Physiology Score (LAPS),[6,10] we developed a predictive model for severity of illness (a continuous variable—the higher the LAPS, the sicker the patient). This model can be conceived as follows:

$$\text{LAPS} = f\,(\text{age, sex, diagnosis, comorbidities, Emergency Department admission or not})$$

This model provided us with coefficients that could be used to assign each non-KPNC hospitalization a simulated severity score,

(Sidebar continued on next page)

(Sidebar continued from previous page)

which we called SIMLAPS. We took additional precaution and assigned the SIMLAPS to all KPNC hospitalizations using the same process. Thus, in our simulation, we explored the effect on the outcomes rates of having the additional information to more accurately measure severity of illness in the non-KPNC hospitals.

RESULTS

Our findings are detailed in the online Appendix for interested readers (available at www.thepermanentejournal.org/files/2017/16-084-Appendix.pdf). They are summarized as follows. Appendix: Part 4 shows that KPNC hospitals are larger and serve larger numbers of Medicare beneficiaries than do most other California hospitals. Most KPNC hospitals performed better than the remaining California hospitals did with respect to both process and outcomes measures. Furthermore, this relationship held when hospitals were assessed globally (ie, all ten conditions pooled together) as well as with respect to individual primary conditions. However, despite being part of an integrated system, residual variation in both process and outcomes measures persists in the system, as is described here and in Appendix: Part 4.

Examination of hospital rankings using the simulated scores did not yield any surprises. Figure 1, which reports data for 3 of the abovementioned 10 conditions, provides an illustrative example. (Additional data are provided in the online Appendix: Part 5.) The overall distribution of rankings showed a similar picture to that obtained without the SIMLAPS for all 3 conditions. However, the range of hospital rankings increased because of the additional error incurred in the SIMLAPS estimates. The KPNC hospitals had better performance, and significant residual variation persisted after incorporation of severity of illness, consistent with what we had observed in our internal severity-adjusted analyses for almost 10 years. Intriguingly, Figure 1 shows that KPNC's performance was best among patients with sepsis, as is evident by the distribution of adjusted mortality rates. Patients with sepsis account for the largest proportion of hospital deaths in both the KPNC (27%) and CMS (30%) cohorts. At KPNC, sepsis care has been the focus of intense analytic and targeted quality improvement efforts since 2008.[11,12]

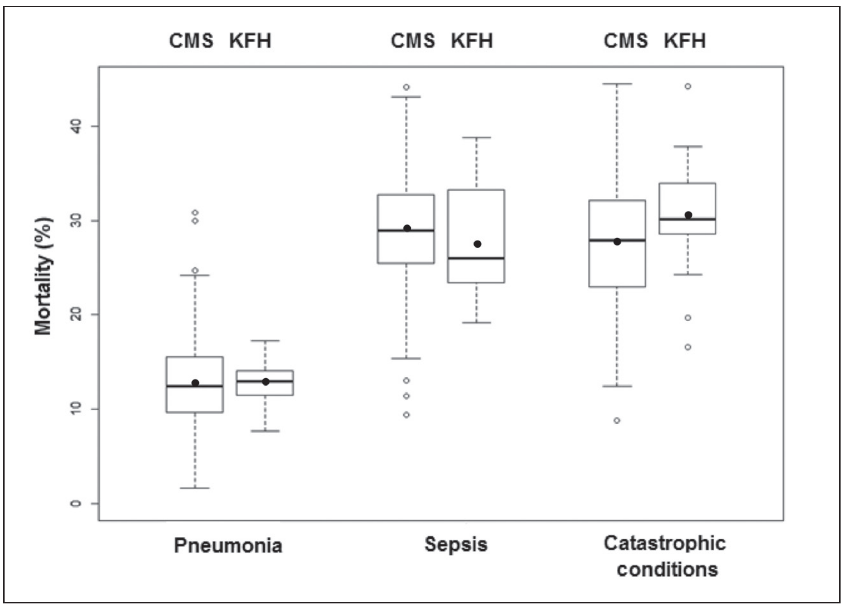

Figure 1. Risk-adjusted 30-day mortality rates for 3 primary conditions with high mortality rates (community-acquired pneumonia, sepsis, and catastrophic conditions) in California hospitals caring for fee-for-service (left, CMS [Centers for Medicare and Medicaid Services]) and Kaiser Foundation Hospitals, Inc (right, KFH) Medicare Advantage beneficiaries.[a]

[a] Analyses control for age, sex, admission venue, principal diagnosis, present-on-admission comorbidities, and a simulated severity of illness score (see text for details). The unit of analysis is a hospital with at least 50 cases. Central dot is the mean mortality rate for all hospitals; boxplot shows the median, interquartile range, and 2.5th and 97.5th percentiles of the mortality distribution across all hospitals; and hollow dots show individual observations outside the 2.5th and 97.5th percentiles.

Figure 2 shows how having SIMLAPS also allowed us to assess the impact of randomly varying LAPS, thus simulating both natural variation (eg, one influenza season might be worse than another) and systemic trends outside KPNC. We did this by randomly shifting the overall severity of illness distribution in the non-KPNC hospitals by 0.15 standard deviation to the left (ie, making non-KPNC patients healthier at the time of admission) or to the right (making them sicker). This is shown graphically in Figure 2, which compares global performance regarding mortality in 3 scenarios: when severity is not included, when it is included but the severity distribution is unaltered, and when the CMS severity distribution is decreased or increased.

Figure 3 shows 2 KPNC quarterly trend lines for the period from March 2010 through May 2015. The dashed line shows the discharge rate from our acute care hospitals, which fell from approximately 72/1000 to approximately 61/1000 members (a 15% decline) during this period. During this time, KPNC made major increases in its hospitalization prevention and case management programs. The solid line shows the average severity of illness of patients admitted to the hospital, using our more sophisticated LAPS Version 2, or LAPS2, which also includes vital signs and patients' neurologic status.[8] During this period, the mean LAPS2 for all admissions increased by 33% (from 54 to 72), whereas other internal analyses have found that the proportion of extremely ill patients (LAPS2 ≥ 110 as well as a high comorbidity burden) almost doubled, from 5% of all emergency admissions to 9%. Thus—some would consider it ironic—success in one area (achieved by enhancing outpatient care and case management efforts) may be leading to problems in another (emergency admissions are sicker, making hospital care for nonelective admissions harder and more complicated). Moreover, it is also clear that KPNC has not eliminated physician-level variation; for example, using very recent internal data, we found that the mean (76 to 100) and median (68 to 99) LAPS2 for patients hospitalized with community-acquired pneumonia both vary considerably across our 21 hospitals. ❖

References

1. Krumholz HM, Normand SL, Bratzler DW, et al. Risk-adjustment methodology for hospital monitoring/surveillance and public reporting. Supplement #1: 30-day mortality model for

(Sidebar continued on next page)

(Sidebar continued from previous page)

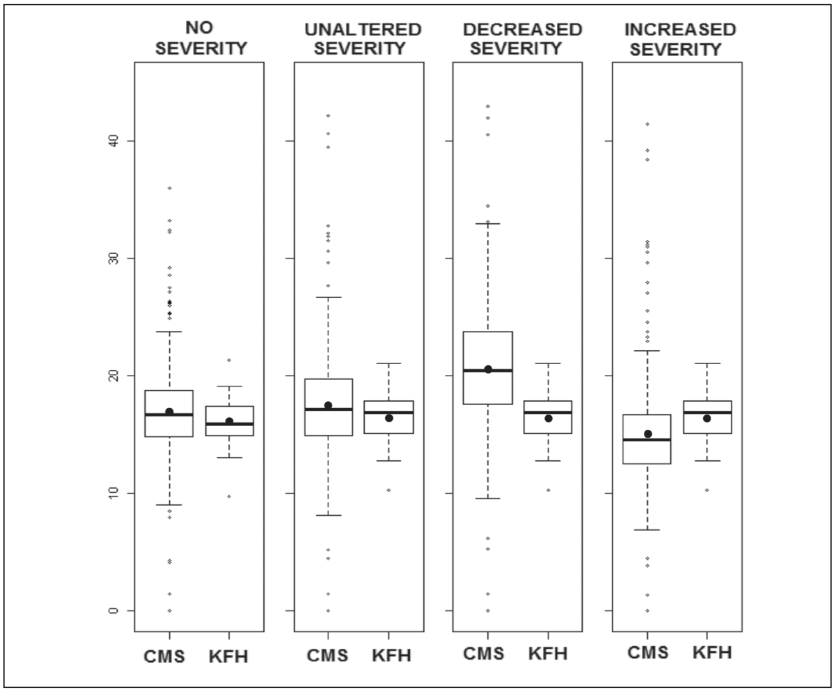

Figure 2. Risk-adjusted 30-day mortality rates (%, y axis) among California hospitals caring for fee-for-service Medicare (left, CMS [Centers for Medicare and Medicaid Services]) and Kaiser Foundation Hospitals (right, KFH) Medicare Advantage beneficiaries.[a]

[a] The unit of analysis is a hospital with at least 50 cases. Central dot is the mean mortality rate for all hospitals; boxplot shows the median, interquartile range, and 2.5th and 97.5th percentiles of the mortality distribution across all hospitals; and dots show individual observations outside the 2.5th and 97.5th percentiles. All analyses control for age, sex, admission venue, principal diagnosis, and present on admission comorbidities. In the first pair of rankings (far left, labeled *NO SEVERITY*) no other variables are included in the risk adjustment, whereas in the second pair (*UNALTERED SEVERITY*), a simulated severity of illness score is added to all hospitalizations. The third pair (*DECREASED SEVERITY)* shows the effect of *decreasing* severity of illness among CMS hospitalizations, whereas the fourth panel (*INCREASED SEVERITY*) shows the effect of *increasing* severity of illness among CMS hospitalizations. See text for description of how severity distributions were varied.

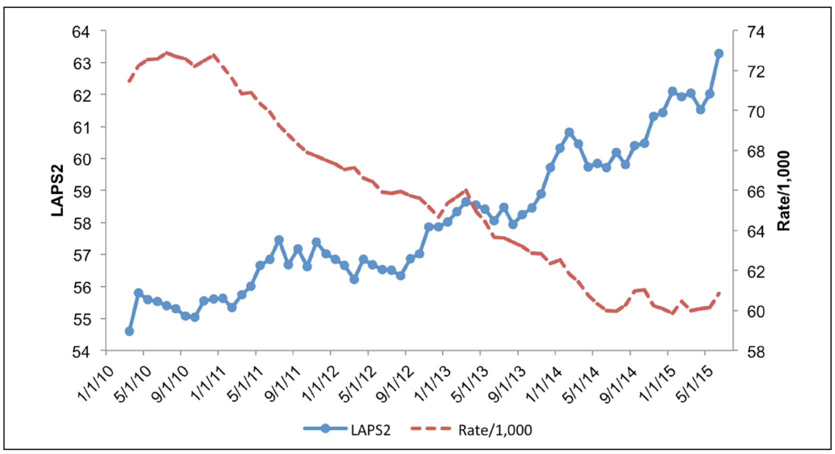

Figure 3. Change in discharge rate and severity of illness.[a]

[a] The figure shows a progressive decrease in the discharge rate of Kaiser Permanente Northern California's 21 acute care hospitals. Concurrently, the average severity of illness score among patients admitted to these hospitals, as measured by the Laboratory-based Acute Physiology Score, version 2 (LAPS2, described in citation 8) has shown a progressive increase over this 5-year time period.

pneumonia [Internet]. Baltimore, MD: Centers for Medicare & Medicaid Services; 2006 [cited 2017 May 16]. Available from: www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228861744769&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DYaleCMS_PN_Report%2C0.pdf&blobcol=urldata&blobtable=MungoBlobs.

2. Krumholz HM, Normand SL, Galusha DH, et al. Risk-adjustment models for AMI and HF 30-day mortality: Methodology [Internet]. Baltimore, MD: Centers for Medicare & Medicaid Services; 2005 [cited 2017 May 16]. Available from: www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228861777994&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DYale_AMI-HF_Report_7-13-05%2C0.pdf&blobcol=urldata&blobtable=MungoBlobs.
3. Bradley EH, Holmboe ES, Mattera JA, Roumanis SA, Radford MJ, Krumholz HM. A qualitative study of increasing beta-blocker use after myocardial infarction: Why do some hospitals succeed? JAMA 2001 May 23-30;285(20):2604-11. DOI: https://doi.org/10.1001/jama.285.20.2604.
4. Bradley EH, Curry LA, Spatz ES, et al. Hospital strategies for reducing risk-standardized mortality rates in acute myocardial infarction. Ann Intern Med 2012 May 1;156(9):618-26. DOI: https://doi.org/10.7326/0003-4819-156-9-201205010-00003.
5. Escobar GJ, Fireman BH, Palen TE, et al. Risk adjusting community-acquired pneumonia hospital outcomes using automated databases. Am J Manag Care 2008 Mar;14(3):158-66.
6. Escobar GJ, Greene JD, Scheirer P, Gardner MN, Draper D, Kipnis P. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. Med Care 2008 Mar;46(3):232-9. DOI: https://doi.org/10.1097/MLR.0b013e3181589bb6.
7. Liu V, Turk BJ, Ragins AI, Kipnis P, Escobar GJ. An electronic simplified acute physiology score-based risk adjustment score for critical illness in an integrated healthcare system. Crit Care Med 2013 Jan;41(1):41-8. DOI: https://doi.org/10.1097/ccm.0b013e318267636e.
8. Escobar GJ, Gardner MN, Greene JD, Draper D, Kipnis P. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated healthcare delivery system. Med Care 2013 May;51(5):446-53. DOI: https://doi.org/10.1097/mlr.0b013e3182881c8e.
9. Render ML, Kim HM, Welsh DE, et al; VA ICU Project (VIP) Investigators. Automated intensive care unit risk adjustment: Results from a National Veterans Affairs study. Crit Care Med 2003 Jun;31(6):1638-46. DOI: https://doi.org/10.1097/01.ccm.0000055372.08235.09.
10. van Walraven C, Escobar GJ, Greene JD, Forster AJ. The Kaiser Permanente inpatient risk adjustment methodology was valid in an external patient population. J Clin Epidemiol 2010 Jul;63(7):798-803. DOI: https://doi.org/10.1016/j.jclinepi.2009.08.020.
11. Whippy A, Skeath M, Crawford B, et al. Kaiser Permanente's performance improvement system, part 3: Multisite improvements in care for patients with sepsis. Jt Comm J Qual Patient Saf 2011 Nov;37(11):483-93. DOI: https://doi.org/10.1016/s1553-7250(11)37061-4.
12. Liu V, Escobar GJ, Greene JD, et al. Hospital deaths in patients with sepsis from 2 independent cohorts. JAMA 2014 Jul 2;312(1):90-2. DOI: https://doi.org/10.1001/jama.2014.5804

is that the severity of illness distribution for admissions coming through its Emergency Department will be shifted (ie, patients will be sicker). Put differently, as outpatient illness prevention improves, the patients coming into the Emergency Department will be more likely to consist of sicker patients in whom preventive efforts failed. The second factor is the admission threshold for individual physicians. It is known that this threshold varies considerably across individuals, as can be seen, for example, with the decision to admit a patient with pneumonia.[27] It is also suspected, although a more mechanistic description still eludes us, that clinicians who practice together tend to practice similarly (ie, admission thresholds will vary *across* hospitals more than *within* hospitals). Thus, aggregate practice variation can affect estimates of the quality of hospital care.[28]

Considering all these factors together, a major limitation of hospital benchmarking became apparent to us; given the potentially powerful effects of admission thresholds, hospital quality of care (as currently measured using only data from *hospitalized* patients) may play a smaller role in explaining variation in hospital outcomes than is assumed. It is entirely possible that hospitals with low thresholds for admission might "benchmark better" than those that do not. This phenomenon would be enhanced if patient case mix measures (particularly those with limited or no severity component) were biased in such a way that less sick patients look sicker than they really are.

The notion that it might not be possible to adjust for case mix underlies a different approach to measuring hospital quality, the concept of "failure to rescue" developed by Silber et al,[29,30] who argue that a hospital's ability to "rescue" a patient after a complication is a better reflection of its quality than its risk-adjusted mortality rate. However, to our knowledge, so far no one has examined the relationship between admission thresholds and "failure to rescue." Furthermore, although the "failure to rescue" construct has an attractive theoretical basis as well as strong face validity among clinicians, it is not being used for routine benchmarking.

Consideration of the importance of the admission threshold has led us to reflect on another factor that can affect hospital rankings: The impact of patients near the end of life. Intuitively, it would seem that, if a hospital were more or less likely to admit patients near the end of life (as opposed to, say, diverting them to hospice), it might have a higher or lower death rate. Thus, the apparent performance of hospitals that admitted more patients near the end of life might appear worse.

Until recently, concerns about this issue have been somewhat theoretical because obtaining information about advance health care directives (eg, a patient's preferences in the event s/he were to experience cardiac arrest in the hospital) on a large scale from paper charts has been difficult.[31] However, in an era in which a large proportion of US hospitals have deployed or are deploying comprehensive EMRs, ignoring this issue will become less tenable.

In previously published work, using data from the Epic inpatient EMR (in which specifying a patient's resuscitation preference is a "hard stop," without which it is not possible to admit a patient), we have found that, despite the limitations of our data systems, one can strongly infer that consideration of patients near the end of life must become an important benchmarking component.[21,32] We reported that approximately 11% of all KPNC hospitalized adults have a "do not resuscitate" order on admission, with another approximately 2% with a "partial code" or "comfort care only" order. Among patients admitted through the Emergency Department, approximately 18% have a "do not resuscitate," "partial code," or "comfort care only" order at admission. Moreover, our published analyses found that the impact of including care directives on hospital rankings is profound. Although our sample included only 21 hospitals, 3 (14%) of the 21 had a statistically significant change in their observed-to-expected mortality ratio when care directives were included (ie, their rankings changed dramatically). This proportion is sobering, given the importance accorded to hospital rankings for public reporting. Furthermore, in 2009, in internal, unpublished analyses in which we employed a 6-month mortality risk measure, we found an almost 3-fold variation across our hospitals in the proportion of patients with a predicted mortality risk of 30% or greater. The problem of face validity cannot be ignored, either, because clinicians may be skeptical of risk adjustment models that do not consider patient physiology or end-of-life care preferences.

WHAT MIGHT FUTURE HOSPITAL BENCHMARKING LOOK LIKE?

One aspect of this is very clear; future risk adjustment methods should incorporate laboratory data, vital signs, nurse-captured indicators (eg, mental status, functional status, care order status, and indicators of frailty), care directives, and oxygenation status. Including these data elements, in addition to improving the quality of the risk adjustment, would enhance benchmarking's face validity among clinicians and health care organizations. In addition, the *range* of outcomes and process measures needs to be expanded; at a bare minimum, benchmarking should include the use of intensive care and assisted ventilation.

However, further research also must be conducted on the admission threshold, including what quantifiable factors determine it and what its impact is when one varies it systematically. This may need to include simulation studies. Some of these studies may need to incorporate more "upstream" data and consider the relationship between inferences made when one varies the unit of analysis. One approach to the analysis of health care processes is to change the unit of analysis from an individual hospital encounter to an episode.[33] Thus, if one wants to assess pneumonia care, the analytic record would not be a pneumonia admission from July 12, 2012, to July 18, 2012, but, rather, one that began on July 9, 2012 (outpatient visit for cough and mild fever), spanned the hospitalization, and included the postdischarge follow-up visit on July 27, 2012. We have employed this episode-based approach to analyze the characteristics of respiratory syncytial virus infections in infants,[34] and KPNC uses a software package that subdivides our population into *episode treatment groups*[35,36] for internal quality assurance and quality improvement. However, if the goal is to assess *hospitals*, this basic approach would need modifications.

We conclude this report with some informed speculation on what kind of research one might conduct that incorporates consideration of the admission threshold (which is driven by events in the *outpatient* setting) in the assessment of *hospital* performance. One important component of this kind of work is the need to go "upstream" and incorporate data on patient status preceding hospitalization. This could take the form of including trending terms for patients' illness severity (eg, incorporate a severity score for the week preceding admission), which are known to increase statistical models' predictive performance.[37] Such work could also incorporate measures identifying whether some sort of screening was taking place for the most common conditions that tend to drive a hospital's overall performance. For example, although we and others have documented that patients with sepsis have high mortality and morbidity *after* their hospitalization, very little work has been done on what happened to such patients *before* hospitalization. Intuitively, patient trajectories as measured by data elements other than severity scores would seem to have high information value. For example, patients with infection who go on to experience sepsis without antibiotic treatment may have very different outcomes from patients with similar illnesses seen in the outpatient clinic and treated with oral antibiotics (ie, such patients, in whom sepsis develops *despite* treatment, may have "hidden" illness severity).

Future studies should also include incorporation of hierarchical variables. These are variables that do not vary by patient, but by hospital. One could test variables that capture hospitals' historical tendency to admit low-risk as well as high-risk patients (including patients near the end of life), for example, the percentage of patients with mortality risk less than 2% or more than 30%, respectively, averaged during a 3-year period. It would also be possible to incorporate variables that measure integration (eg, proportion of patients belonging to health plans with hospitalization prevention systems in place). A variety of methods, including simulation, would need to be employed, and it would make sense to make systematic comparisons of rankings when models do and do not include such hierarchical variables. These kinds of analyses would also require a larger sample of hospitals.

Finally, it is imperative that we expand the domain of measures that address how hospitals respect patient choices regarding care near the end of life. Although incorporation of care directive data in risk adjustment models is important, it can only be considered a first step, particularly given the fact that, with respect to hospital care of such patients, the critical component is the decision to admit at all.

CONCLUSION

Practice variation undermines medicine's moral authority and the notion that medicine is a science rather than an art. In the face of analyses that control for patient characteristics but still show variation in processes, costs, and outcomes, the presence of residual variation raises not just questions of fairness and competence[38-40] but even of the very basis for valuing health care. This thought is summarized by two scholars, Stuart H Altman and Uwe E Reinhardt, as follows[41]: "[S]ignificant variations in the per capita use of health care, unrelated to differences in outcomes, undermined the traditional argument that reductions in health care spending would inevitably entail commensurate reductions in the quality of health care."

Benchmarking is a critical tool in the struggle against unnecessary practice variation. Properly conducted—and when situated in receptive health systems—it can motivate substantial quality improvement and thus save lives, improve quality, and decrease suffering. In addition to conducting research on improving actual hospital benchmarking, our profession must also do a better job of explaining it to the public, which should include informed "blue-sky" speculation and use of simulation. After all, other scientists share their speculations all the time. Isn't it time that health services researchers did so as well? ❖

Disclosure Statement

The author(s) have no conflicts of interest to disclose.

Acknowledgments

This work was supported by The Permanente Medical Group, Inc, and Kaiser Foundation Hospitals, Inc, and was approved by the Kaiser Permanente Northern California Institutional Review Board for the Protection of Human Subjects. None of the sponsors had any involvement in our decision to submit this manuscript or in the determination of its contents. Dr Liu was supported by the National Institute of General Medical Sciences Award K23GM112018 from the National Institutes of Health, Bethesda, MD.

We thank Laurence Baker, PhD, for assistance in understanding the process of acquisition of data from the Centers for Medicare and Medicaid Services (CMS). We are very grateful to the Yale Center for Outcomes Research and Evaluation (Harlan Krumholz, MD; Susannah Bernheim, MD, MHS; and Jackie Grady, MS) in New Haven, CT, for their assistance with understanding the CMS database structure. We thank Philip Madvig, MD; Cesar Villalpando; and Kathy Weiner for their administrative support; Tracy Lieu, MD, MPH, for reviewing the manuscript; and Rachel Lesser and Vanessa Floyd-Rodriguez, MPH, for assistance with formatting and editing.

Kathleen Louden, ELS, of Louden Health Communications provided editorial assistance.

How to Cite this Article

Escobar GJ, Baker JM, Turk BJ, Draper D, Liu V, Kipnis P. Comparing hospital processes and outcomes in California Medicare beneficiaries: Simulation prompts reconsideration. Perm J 2017;21:16-084. DOI: https://doi.org/10.7812/TPP/16-084.

References

1. Avery ME, Tooley WH, Keller JB, et al. Is chronic lung disease in low birth weight infants preventable? A survey of eight centers. Pediatrics 1987 Jan;79(1):26-30.
2. Krumholz HM, Normand SL, Spertus JA, Shahian DM, Bradley EH. Measuring performance for treating heart attacks and heart failure: The case for outcomes measurement. Health Aff (Millwood) 2007 Jan -Feb;26(1)75-85. DOI: https://doi.org/10.1377/hlthaff.26.1.75.
3. Webster TR, Curry L, Berg D, Radford M, Krumholz HM, Bradley EH. Organizational resiliency: How top-performing hospitals respond to setbacks in improving quality of cardiac care. J Healthc Manag 2008 May-Jun;53(3):169-81.
4. Merle V, Moret L, Pidhorz L, et al. Does comparison of performance lead to better care? A pilot observational study in patients admitted for hip fracture in three French public hospitals. Int J Qual Health Care 2009 Oct;21(5):321-9. DOI: https://doi.org/10.1093/intqhc/mzp029.
5. Austin JM, Jha AK, Romano PS, et al. National hospital ratings systems share few common scores and may generate confusion instead of clarity. Health Aff (Millwood) 2015 Mar;34(3):423-30. DOI: https://doi.org/10.1377/hlthaff.2014.0201.
6. Wachter RM. How measurement fails doctors and teachers [Internet]. New York, NY: The New York Times Sunday Review; 2016 Jan 16 [cited 2016 Jan 22]. Available from: www.nytimes.com/2016/01/17/opinion/sunday/how-measurement-fails-doctors-and-teachers.html?_r=0.
7. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: The problem with small sample size. JAMA 2004 Aug 18;292(7):847-51. DOI: https://doi.org/10.1001/jama.292.7.847.
8. Lilford R, Pronovost P. Using hospital mortality rates to judge hospital performance: A bad idea that just won't go away. BMJ 2010 Apr 20;340:c2016. DOI: https://doi.org/10.1136/bmj.c2016.

9. Harder B, Comarow A. Hospital quality reporting by US News & World Report: Why, how, and what's ahead. JAMA 2015 May 19;313(19):1903-4. DOI: https://doi.org/10.1001/jama.2015.4566.
10. Comarow A. An opt-in program that lets health systems supplement Medicare claims data [Internet]. Washington, DC: U.S. News & World Report; 2015 May 7 [cited 2015 Oct 22]. Available from: http://health.usnews.com/health-news/blogs/second-opinion/2015/05/07/an-opt-in-program-that-lets-health-systems-supplement-medicare-claims-data.
11. Cram P, Cai X, Lu X, Vaughan-Sarrazin MS, Miller BJ. Total knee arthroplasty outcomes in top-ranked and non-top-ranked orthopedic hospitals: An analysis of Medicare administrative data. Mayo Clin Proc 2012 Apr;87(4):341-8. DOI: https://doi.org/10.1016/j.mayocp.2011.11.017.
12. Medicare.gov. Hospital compare [Internet]. Baltimore, MD: US Centers for Medicare & Medicaid Services; 2017 [cited 2017 Apr 27]. Available from: www.medicare.gov/hospitalcompare/.
13. Jacobson G, Damico A, Neuman T, Gold M. Medicare Advantage 2015 spotlight: Enrollment market update [Internet]. Menlo Park, CA: Henry J. Kaiser Family Foundation; 2015 Jun [cited 2017 May 1]. Available from: http://files.kff.org/attachment/issue-brief-medicare-advantage-2015-spotlight-enrollment-market-update.
14. Selby JV. Linking automated databases for research in managed care settings. Ann Intern Med 1997 Oct 15;127(8 Pt 2):719-24. DOI: https://doi.org/10.7326/0003-4819-127-8_part_2-199710151-00056.
15. US News & World Report announces the 2016 best Medicare plans [Internet]. Washington, DC: US News & World Report; 2015 [cited 2015 Oct 22]. Available from: www.usnews.com/info/blogs/press-room/2015/10/15/us-news-announces-the-2016-best-medicare-plans.
16. CMS.gov. Five-star quality rating system [Internet]. Baltimore, MD: Centers for Medicare & Medicaid Services; updated 2017 Apr 26 [cited 2015 Oct 22]. Available from: www.cms.gov/medicare/provider-enrollment-and-certification/certificationandcomplianc/fsqrs.html.
17. Escobar GJ, Greene JD, Scheirer P, Gardner MN, Draper D, Kipnis P. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. Med Care 2008 Mar;46(3):232-9. DOI: https://doi.org/10.1097/MLR.0b013e3181589bb6.
18. Liu V, Kipnis P, Gould MK, Escobar GJ. Length of stay predictions: Improvements through the use of automated laboratory and comorbidity variables. Med Care 2010 Aug;48(8):739-44. DOI: https://doi.org/10.1097/mlr.0b013e3181e359f3.
19. Escobar GJ, Greene JD, Gardner MN, Marelich GP, Quick B, Kipnis P. Intra-hospital transfers to a higher level of care: Contribution to total hospital and intensive care unit (ICU) mortality and length of stay (LOS). J Hosp Med 2011 Feb;6(2):74-80. DOI: https://doi.org/10.1002/jhm.817.
20. Liu V, Turk BJ, Ragins AI, Kipnis P, Escobar GJ. An electronic simplified acute physiology score-based risk adjustment score for critical illness in an integrated healthcare system. Crit Care Med 2013 Jan;41(1):41-8. DOI: https://doi.org/10.1097/ccm.0b013e318267636e.
21. Escobar GJ, Gardner MN, Greene JD, Draper D, Kipnis P. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated healthcare delivery system. Med Care 2013 May;51(5):446-53. DOI: https://doi.org/10.1097/mlr.0b013e3182881c8e.
22. Ballesca MA, LaGuardia JC, Lee PC, et al. An electronic order set for acute myocardial infarction is associated with improved patient outcomes through better adherence to clinical practice guidelines. J Hosp Med 2014 Mar;9(3):155-61. DOI: https://doi.org/10.1002/jhm.2149.
23. Pine M, Jordan HS, Elixhauser A, et al. Enhancement of claims data to improve risk adjustment of hospital mortality. JAMA 2007 Jan 3;297(1):71-6. DOI: https://doi.org/10.1001/jama.297.1.71.
24. Tabak YP, Johannes RS, Silber JH. Using automated clinical data for risk adjustment: Development and validation of six disease-specific mortality predictive models for pay-for-performance. Med Care 2007 Aug;45(8):789-805. DOI: https://doi.org/10.1097/mlr.0b013e31803d3b41.
25. Render ML, Kim HM, Welsh DE, et al; VA ICU Project (VIP) Investigators. Automated intensive care unit risk adjustment: Results from a National Veterans Affairs study. Crit Care Med 2003 Jun;31(6):1638-46. DOI: https://doi.org/10.1097/01.ccm.0000055372.08235.09.
26. Ezekowitz JA, Kaul P, Bakal JA, Quan H, McAlister FA. Trends in heart failure care: Has the incident diagnosis of heart failure shifted from the hospital to the emergency department and outpatient clinics? Eur J Heart Fail 2011 Feb;13(2):142-7. DOI: https://doi.org/10.1093/eurjhf/hfq185.
27. Dean NC, Jones JP, Aronsky D, et al. Hospital admission decision for patients with community-acquired pneumonia: Variability among physicians in an emergency department. Ann Emerg Med 2012 Jan;59(1):35-41. DOI: https://doi.org/10.1016/j.annemergmed.2011.07.032.
28. Miller MG, Miller LS, Fireman B, Black SB. Variation in practice for discretionary admissions. Impact on estimates of quality of hospital care. JAMA 1994 May 18;271(19):1493-8. DOI: https://doi.org/10.1001/jama.1994.03510430047033.
29. Silber JH, Williams SV, Krakauer H, Schwartz JS. Hospital and patient characteristics associated with death after surgery. A study of adverse occurrence and failure to rescue. Med Care 1992 Jul;30(7):615-29. DOI: https://doi.org/10.1097/00005650-199207000-00004.
30. Silber JH, Rosenbaum PR, Williams SV, Ross RN, Schwartz JS. The relationship between choice of outcome measure and hospital rank in general surgical procedures: Implications for quality assessment. Int J Qual Health Care 1997 Jun;9(3):193-200. DOI: https://doi.org/10.1093/intqhc/9.3.193.
31. Tabak YP, Johannes RS, Silber JH, Kurtz SG. Should do-not-resuscitate status be included as a mortality risk adjustor? The impact of DNR variations on performance reporting. Med Care 2005 Jul;43(7):658-66. DOI: https://doi.org/10.1097/01.mlr.0000167106.09265.4e.
32. Kim YS, Escobar GJ, Halpern SD, Greene JD, Kipnis P, Liu V. The natural history of changes in preferences for life-sustaining treatments and implications for inpatient mortality in younger and older hospitalized adults. J Am Geriatr Soc 2016 May;64(5):981-9. DOI: https://doi.org/10.1111/jgs.14048.
33. Iezzoni L, editor. Risk adjustment for measuring healthcare outcomes. 4th ed. Chicago, IL: Health Administration Press; 2013.
34. Flaherman VJ, Ragins AI, Li SX, Kipnis P, Masaquel A, Escobar GJ. Frequency, duration and predictors of bronchiolitis episodes of care among infants ≥ 32 weeks gestation in a large integrated healthcare system: A retrospective cohort study. BMC Health Serv Res 2012 Jun 8;12:144. DOI: https://doi.org/10.1186/1472-6963-12-144.
35. Forthman MT, Dove HG, Wooster LD. Episode Treatment Groups (ETGs): A patient classification system for measuring outcomes performance by episode of illness. Top Health Inf Manage 2000 Nov;21(2):51-61.
36. Optum. Learn about ETGs [Internet]. Eden Prairie, MN: Optum; 2015 [cited 2015 Oct 11]. Available from: https://etg.optum.com/etg-links/learn-about-etgs/.
37. Kuzniewicz M, Draper D, Escobar GJ. Incorporation of physiological trend and interaction effects in neonatal severity of illness scores: An experiment using a variant of the Richardson score. Intensive Care Med 2007 Sep;33(9):1602-8. DOI: https://doi.org/10.1007/s00134-007-0714-z.
38. Blumenthal D. The variation phenomenon in 1994. N Engl J Med 1994 Oct 13;331(15):1017-8. DOI: https://doi.org/10.1056/nejm199410133311511.
39. Blumenthal D. Quality of care—what is it? N Engl J Med 1996 Sep 19;335(12):891-4. DOI: https://doi.org/10.1056/NEJM199609193351213.
40. Blumenthal D. The origins of the quality-of-care debate. N Engl J Med 1996 Oct 10;335(15):1146-9. DOI: https://doi.org/10.1056/NEJM199610103351511.
41. Altman SH, Reinhardt UE. Where does health care reform go from here? An unchartered odyssey. Baxter Health Policy Rev 1996;2:xxi-xxxii.

Convalescence

The sooner patients can be removed from the depressing influence of general hospital life the more rapid their convalescence.

— Charles H Mayo, 1865-1939, American medical practitioner and one of the founders of the Mayo Clinic