Brooklyn Law Review

Volume 58 | Issue 1 Article 4

1-1-1992

Problems in the Use of Outcome Statistics to Compare Health Care Providers

Jesse Green

PROBLEMS IN THE USE OF OUTCOME STATISTICS TO COMPARE HEALTH CARE PROVIDERS

*Jesse Green**

INTRODUCTION

Until just a few years ago, disclosure of the names of hospitals or physicians along with their patients' death rates would have been considered indiscreet. In 1986 the Health Care Financing Administration ("HCFA") made the first public release of such a list, but only after the *New York Times* obtained the data on appeal under the federal Freedom of Information Act (FOIA).[1]

While initially reticent, HCFA and others in possession of such data have subsequently adapted to the new openness. Since 1987 HCFA has held annual news conferences to release mortality rates of all United States hospitals treating Medicare patients. Press kits distributed at these sessions include the widely publicized "death list" of high mortality outlier hospitals. This ranking is instantly transmitted by wire services to television newsrooms and newspapers throughout the country. The following headlines appearing on May 2, 1991 indicate the range of coverage: "161 Hospitals Found With High Death Rates";[2] "Seven Hospitals in Pennsylvania Cited for Deaths";[3] and "City's 'Killer' Hospitals Listed."[4]

Since HCFA's initial public release of this information, interest in the use of comparative outcome statistics to evaluate health care providers has mushroomed.[5] HCFA's efforts spurred

---

* Ph.D. and Director of the Department of Health Policy Research, New York University Medical Center/NYU School of Medicine.

[1] Joel Brinkley, *U.S. Releasing Lists of Hospitals With Abnormal Mortality Rates*, N.Y. TIMES, March 12, 1986, at A1.

[2] *161 Hospitals Found with High Death Rates*, N.Y. TIMES, May 2, 1991, at B11.

[3] *Seven Hospitals in Pennsylvania Cited for Deaths*, LANCASTER NEWS, May 2, 1991, at 4.

[4] Penny Bender, *City's "Killer" Hospitals Listed*, N.Y. DAILY NEWS, May 2, 1991, at 2.

[5] Arnold M. Epstein, M.D., M.A., *The Outcomes Movement—Will It Get Us Where*

a major research initiative focused on effectiveness of care[6] and inspired similar analyses by states,[7] private payers,[8] providers[9] and hospital associations.[10] Minnesota's Blue Cross and Blue Shield has carried the approach beyond merely issuing reports. It now adjusts hospital reimbursement rates depending on statistically measured outcomes to financially reward or punish hospitals for their "performance."[11] New York State broke new ground by publishing post-surgical patient mortality rates for all its cardiac surgeons.[12]

This important trend of releasing performance outcomes of doctors and hospitals raises the controversial issue of whether the results of such studies, insofar as they purport to provide information useful for consumers, should be believed. It seemed timely, therefore, to examine the scientific basis upon which such studies rest and to discuss some of the concerns about their validity. Evaluation of the information contained in lists of provider-specific outcome indicators involves two central issues. First is the problem of "bias," that is, the fact that the outcomes being compared reflect not only the care provided but the condition of the patients each provider served. Second is the issue of random variation, particularly the so-called problem of "multi-

---

We Want To Go?, 323 NEW ENG. J. MED. 266 (1990); Steven F. Jencks, M.D., M.P.H, *Quality Assurance*, 263 JAMA 2679 (1990).

[6] Omnibus Budget Reconciliation Act of 1989, Pub. L. No. 101-239, 103 Stat. 2228 (codified as amended in scattered sections of 42 U.S.C.); Arnold S. Relman, M.D., *Assessment and Accountability*, 319 NEW ENG. J. MED. 1220 (1988); William L. Roper, M.D. et al., *Effectiveness in Health Care*, 319 NEW ENG. J. MED. 1197 (1988); AHCPR Program Note, *Medical Treatment Effectiveness Research*, U.S. Dept. of Health and Human Services; Rockville, MD, March 1990.

[7] PA. HEALTH CARE COST CONTAINMENT COUNCIL, HOSPITAL EFFECTIVENESS REPORT, Pub. No. HE6-1-89, (1989); Edward L. Hannan, Ph.D. et al., *Adult Open Heart Surgery in New York State*, 264 JAMA 2768 (1990).

[8] Purchasers will use outcome data to select quality hospitals. Hospital Peer Review § 14, at 124-26 (1989).

[9] U.S. GENERAL ACCOUNTING OFFICE, VA HOSPITAL CARE: A COMPARISON OF VA AND HCFA METHODS FOR ANALYZING PATIENT OUTCOMES, GAO/PEMD-88-29, Washington, DC, June, 1988.

[10] Quality Measurement and Management Project, The Hospital and Research Educational Trust, Risk-adjusted 30-day Mortality of Fresh Acute Myocardial Infarctions: The User's Guide 1989, *Maryland Hospital Association, Hospital Mortality Data Analysis October 1989* (1989).

[11] M. Darby, Minnesota Blues Re-allocates Resources Using Outcomes Research (1992).

[12] David Zinman, *State Takes Docs' List to Heart*, N.Y. NEWSDAY, Dec. 18, 1991, at 7.

ple comparisons," that is, the fact that statistical theory predicts that when many similar providers are compared, even if no differences exist, some will be "outliers" by chance alone.

These problems are not exclusively the result of any specific shortcomings of particular studies but are to be expected from any study that attempts to compare outcomes of many providers on the basis of events that occurred outside of a controlled experiment. However, in addition to the intrinsic limitations that apply to all such studies, specific features of each outcome study, including the reliability and validity of measured risk factors and analytic techniques employed, also require evaluation in understanding the findings such efforts yield.

I. The Scientific Basis of Comparative Outcome Statistics

A list of hospital or physician-specific mortality rates, adjusted for patient risk, may be viewed by consumers of health services as a basis for comparing the effectiveness of care available from different providers. In assessing the credibility of such data for this purpose, it is useful to consider the more general problem of how medical researchers compare alternative treatments to determine their relative efficacy. What constitutes credible evidence that one therapy—e.g. a drug or surgical intervention—yields better results than another? Under many circumstances the ideal method for generating such evidence is the randomized clinical trial.[13] Generally, in clinical trials groups of patients receive alternative therapies, or one group gets a specific therapy while the other receives a placebo.

An important methodological feature of a randomized clinical trial is that patients are assigned to a group by random selection. Because random assignment is used, any difference in outcomes between groups can come from two sources only: sampling variability and the alternative treatments themselves. The great advantage of this study design is that when randomization is used to assign patients to treatments, the probability that different outcomes are due to sampling variation can be determined precisely using statistical sampling theory. Hence, randomization assures that it is possible to know how confident

[13] David P. Byar, *The Necessity and Justification of Randomized Clinical Trials, in* Controversies in Cancer (HJ Tagnon, M.D. & M.J. Staquet, M.D. eds., 1979).

one may be that observed differences in outcomes are attributable to the treatments being compared.

II. Limitations of Non-Randomized Studies

When patients are assigned non-randomly to different treatments, inferences about the relationship of treatments to outcomes are weakened. When comparing groups that are not random samples of a population, there may be many reasons that could account for observed differences in outcomes, other than alternative treatments. For example, patients given treatment A may have been more seriously ill than those given treatment B. In some situations, the fact that they were more ill may be the reason for prescribing A.[14] If treatment A is thought to be a safer choice in high risk cases, then its use in such cases will be more likely to make the treatment look less safe or less effective than treatment B because high risk cases have higher mortality rates.[15] In other words, considerations of patients' clinical conditions greatly influence treatment choices in practice and are strong potential sources of bias in comparisons of their effects.

Overcoming such biases in non-randomized studies is generally attempted through statistical adjustment. Such adjustment techniques seek to measure and control for the effect of differences between treatment groups, such as age, gender, diagnosis, comorbid conditions and measures of physiologic functioning. This analytic procedure is employed in an effort to simulate mathematically the experimental situation by estimating how different groups would compare if patient characteristics were identical. Unfortunately, while statistical adjustment may reduce bias, it cannot yield that which randomization provides: a quantifiable degree of certainty that the groups are equivalent.

Theoretically, it is possible to conduct a randomized trial of several different health care providers, such as two or more heart surgeons or hospitals. However, such studies are rarely undertaken because of political issues, e.g., patients would be required to accept random assignments and to give up their ability to choose their own surgeons. Instead, observational databases of

---

[14] Sylvia B. Green & David P. Byar, *Using Observational Data from Registries to Compare Treatments*, 3 Stat. in Med. 361, 362 (1984).

[15] Nathan Mantel, *Cautions on the Use of Medical Databases*, 2 Stat. in Med. 355 (1983).

health care episodes are generally employed in the studies that compare providers, such as hospital discharge abstracts, possibly supplemented with additional clinical data.

III. The Role of Chance in Identifying Outliers: The Problem of Multiple Comparisons

Some reports flag providers as "outliers" if their mortality rates are more than two standard deviations higher than predicted. For example, HCFA's report flags hospitals in this manner. This approach treats the question of whether a hospital's mortality rate is greater than predicted as a hypothesis test with a $p < .05$ level of confidence. When many hypothesis tests are conducted in this manner, the problem of multiple inferences arises.[16] As Frederick A. Connell explains, "if the level of significance is set at $p < .05$, 5% of all relationships will be statistically significant by chance alone."[17] This kind of consideration led D.C. Thomas to caution that "[a]ny study which collects information on a large number of 'stimulus' and 'response' variables has a high probability of producing wild goose chases which can consume much valuable research time and resources to refute."[18]

Published reports that present outcome statistics may examine hundreds or thousands of providers, treating each specific provider as a study variable or "risk factor." When considering such lists it is important to note that even under the "null hypothesis"—a scenario in which the providers' performances do not differ at all—5% of providers would be expected to be falsely identified as either high or low "outliers" by a two standard deviation test. In a 1990 report examining data from 5,685 hospitals, HCFA designated 161, or 2.8%, as high mortality outliers.[19] By chance alone one would expect 2.5% of observations, or 142 hospitals, to lie more than two standard deviations above (and a similar number below) the prediction.

---

[16] D.C. Thomas et al., *The Problem of Multiple Inference in Studies Designed to Generate Hypotheses*, 122 Am. J. Epidemiology 1080 (1985).

[17] Frederick A. Connell, *The Use of Large Data Bases in Health Care Studies*, 1987 Ann. Rev. Pub. Health 62.

[18] Thomas, *supra* note 16, at 1081.

[19] *161 Hospitals Found With High Death Rates*, N.Y. Times, May 2, 1991, at B11.

IV. HCFA'S MEDICARE MORTALITY REPORT: A RESEARCH SUMMARY

Death rates among hospitalized Medicare patients at more than 5,000 hospitals have been analyzed by HCFA in its annual series *Medicare Hospital Mortality Information.*[20] In each report, hospitals with observed mortality rates significantly exceeding HCFA's predictions have been flagged as high-mortality outlier hospitals. HCFA's analyses take account of some characteristics of each hospital's patients through a multiple regression model that determines each patient's risk of death as calculated from information contained in Medicare claims records.

To date four published studies have assessed the methodology used in HCFA's mortality report.[21] All four studies indicate that biases were operating in the designation of hospital mortality outliers. In one study, data on the severity of each patient's illness measured on a scale of one to four were used to supplement the basic demographic and diagnostic data employed by HCFA.[22] The study demonstrated that severity of illness was the explanation for most of the gap between actual and predicted mortality rates in a sample of 34,552 patients with 5 diagnoses at 13 hospitals. When the different severity levels were considered, there were distinct subgroups of patients whose risks of death at the time of admission were significantly different. Actual mortality rates among patients in the "cancer" diagnostic category, for example, varied from less than 1% at the lowest severity level to 86% at the highest. Using a method similar to that of HCFA to predict the mortality rates of these patients, the study found that the prediction model was unable to detect

---

[20] OTIS R. BOWEN, M.D. & WILLIAM L. ROPER, M.D., U.S. DEP'T OF HEALTH & HUMAN SERVICES, IV MEDICARE HOSPITAL MORTALITY INFORMATION (1986); OTIS R. BOWEN, M.D. & WILLIAM L. ROPER, M.D., U.S. DEPT OF HEALTH & HUMAN SERVICES, MEDICARE HOSPITAL MORTALITY INFORMATION, REGION II (1987); LOUIS W. SULLIVAN, M.D. & LOUIS B. HAYS, U.S. DEPT. OF HEALTH & HUMAN SERVICES, 5 MEDICARE HOSPITAL MORTALITY INFORMATION (1986-88).

[21] Jesse Green, Ph.D. et al., *Analyzing Hospital Mortality: The Consequences of Diversity in Patient Mix*, 265 JAMA 1849 (1991) [hereinafter *Hospital Mortality*]; Jesse Green, Ph.D. et al., *The Importance of Severity of Illness in Assessing Hospital Mortality*, 263 JAMA 241 (1990) [hereinafter *Severity of Illness*]; David A. Schwartz, M.D., M.P.H. and Philip Reilly, J.D., M.D., *The Choice Not to be Resuscitated*, J. AM. GERIATRICS SOC'Y, Nov. 1986, at 807; David W. Smith, Ph.D. et al., *Using Clinical Variables to Estimate the Risk of Patient Mortality*, 29 MED. CARE 1108 (1991).

[22] Green, *Severity of Illness*, *supra* note 21, at 241.

the differences in severity of illness. In each diagnostic category it predicted mortality rates that did not significantly differ across the severity levels.

David W. Smith and his colleagues compared HCFA's prediction model with a model developed from data that included detailed clinical findings concerning each patient.[23] The authors noted that "the HCFA methods were found to be biased in identifying outlier hospitals" and concluded that some of the bias could be removed if additional clinical data were used to supplement HCFA data.[24]

A RAND Corporation report found that "high mortality outlier" hospitals had as good or better quality of care than non-outlier hospitals and that the higher than predicted mortality rates at the outlier hospitals were due to severity of illness, "do not resuscitate" status and chance variation.[25]

Finally, Jesse Green and his colleagues examined data from 5,560 hospitals and found that patient characteristics among Medicare populations served by different hospitals varied widely and that these differences influenced the mortality results.[26] Even though the HCFA model controls for age, hospitals with the oldest patients were twice as likely to be flagged as mortality outliers as hospitals with the smallest proportion of very elderly patients. An increased percentage of patients requiring nursing home care also resulted in a greater chance of being designated as an outlier hospital.

HCFA has acknowledged that its list of "high mortality outlier hospitals" includes some institutions that provide quality care but have high mortality rates because the patients they serve are severely ill.[27] HCFA estimates that such errors account for one-third of listed hospitals.[28] However, the study by Green and his colleagues indicated that the figure may be as high as 46%.[29]

---

[23] Smith, *supra* note 21, at 1110.

[24] *Id.* at 1116-19.

[25] Rolla E. Park, Ph.D, M.B.A. et al., *Explaining Variations in Hospital Death Rates: Randomness, Severity of Illness, Quality of Care*, 264 JAMA 484 (1990).

[26] *Hospital Mortality, supra* note 21.

[27] David Zinman, *Hospitals' Deaths at Expected Levels*, NEWSDAY, May 2, 1991, at 19.

[28] *Id.*

[29] *Hospital Mortality, supra* note 21, at 1852.

V. The Fallacy of Interpreting Unexplained Differences as Effectiveness

Biases in published mortality statistics are not surprising given the inherent limitations of observational databases for addressing outcome differences. As noted above, similar problems arise when observational databases are used to study treatment efficacy. For example, David P. Byar analyzed thyroid tumor registry data and found that prognostic factors, such as age, sex, histology of tumor, size and location of tumor and extent of metastases, were associated with enormous variation in patients' survival duration. Byar concluded that such data could generate misleading results because ". . . any differences in survival attributable to different therapies are likely to be much smaller than those attributed to prognostic factors."[30]

By the same token, given the substantial variation among providers in terms of the characteristics of patients served and the impact of these characteristics on mortality, the difficulty of detecting any differences that are genuinely due to "quality of care" or effectiveness is problematic. Compounding this problem is that while clinical characteristics of patients have a very large effect on mortality rates, the effect on quality may be quite small. For example, a peer review organization chart audit of 275,274 randomly selected Medicare hospitalizations turned up less than 0.5% involving an avoidable death.[31] Rolla E. Park and her colleagues reported that marked differences in quality of care provided to patients with congestive heart failure were associated with a 1.5% gradient in 30-day mortality.[32]

With each improvement in data or technique there is a temptation to label whatever variation not yet accounted for as "quality of care" or "effectiveness." However, detecting small differences in death rates that would be expected to result from variations in effectiveness may be beyond the power of existing methods. In 1989 HCFA flagged 182 hospitals as high outliers based on patients admitted in 1986. No hospital was flagged unless its actual mortality rate was at least 3.1% above HCFA's prediction and the majority of flagged hospitals had mortality

---

[30] Byar, *supra* note 13, at 76.

[31] Prospective Payment Assessment Commission, Medicare Prospective Payment and the American Health Care System, Report to Congress, June 1988.

[32] Park, *supra* note 25, at 487.

rates at least 6% greater than expected.[33] Allowing such wide margins means, for example, that a hospital with 1,000 cases and an expected mortality rate of 12%, or 120 deaths, would avoid "outlier" designation if its mortality rate was 15%, or 150 deaths, and would probably not be flagged unless its mortality rate rose to 18%, or 180 deaths. "Excess" deaths of such magnitude *could* be accounted for if a hospital provided care to many severely ill patients with "do not resuscitate orders" on admission.[34] However, the view that it could be due to quality of care requires a much higher prevalence of unnecessary or preventable deaths in hospitals than has been reported. This illustrates why the belief that very large differences in mortality rates are likely to be due to quality may be fallacious.

A. *Severity of Illness*

The need to measure and control for disease severity is accentuated when patients with the same diagnosis may have different prognoses depending on their degree of illness. A variety of techniques have been developed for estimating the severity of a patient's illness. Some rely solely on hospital discharge data. For example, HCFA's method used for the annual mortality report projects a prognosis for each patient expressed as the *a priori* probability that despite treatment, the patient would die within each of three fixed time periods—30, 90 and 180 days of admission. These probabilities are calculated from claims data using listed diagnoses, age and sex of the patient, and severity proxies, that is, whether the patient came from home, a nursing home, or elsewhere and whether the "type" of admission was "elective," "urgent" or "emergent."

Other approaches to measuring severity require a much broader array of clinical findings than are available on claims data. For example, MedisGroups, a proprietary computerized system, generates severity of illness scores ranging from one (least severe) to four (most severe) by applying an algorithm to the results of physiologic measures.[35] There are a number of

---

[33] Jesse Green, Ph.D. et al., Hospital Mortality Reports 12 (July 31, 1990) (final report to the AARP Andrus Found. on file with the author).

[34] Park, *supra* note 25, at 487.

[35] Lisa I. Iezzoni, M.D. et al., *Admission and Mid-Stay MedisGroups Scores as Predictors of Death Within 30 Days of Hospital Admission*, 81 AM. J. PUB. HEALTH 74,

published reviews of the available severity measurement systems.[36] In general, the literature supports the view that explicit severity measures add considerably to the predictive capability of mortality prediction models.[37] Nonetheless, biased assessments may still result from flaws in the measurement of severity by some of these instruments.[38]

One purpose of the hospital mortality listings is to indicate the effectiveness of caregiver performance to consumers. In this context, severity of illness measures are used to separate two components of the mortality rate: (1) the degree to which the patient's present condition contributed to the probability of death, and (2) the contribution of the specific care provider. While statistical adjustment cannot be expected to disaggregate thoroughly these two effects, inadequate adjustment leads to mortality statistics that are completely uninterpretable.

Much of the debate surrounding the release of provider-specific mortality data focuses on the adequacy of the statistical adjustment for severity of illness. Because much of this debate can be rather arcane, its importance may be lost in situations where conflicting interests compete, such as in litigation. Accordingly, it would be helpful to have an agreed upon set of criteria that could be used to evaluate the various techniques used to measure and adjust for severity of illness. While some criteria, such as measures of predictive validity, are commonly cited, there is no broad consensus either on which measure is most appropriate, or what level of predictive validity is adequate for a given purpose.

One measure that is frequently used to indicate the predictive power of a severity measure when used to project mortality

---

74-75 (1991).

[36] Farrokh Alemi, Ph.D. et al., *Predicting In-Hospital Survival of Myocardial Infarction*, 28 MED. CARE 762 (1990); David Aquilina et al., *Using Severity Data to Measure Quality*, BUS. & HEALTH, June 1988, at 40; Lisa I. Iezzoni, M.D. et al., *Predicting In-Hospital Mortality*, 30 MED. CARE 347 (1992) [hereinafter *In-Hospital Mortality*]; Lisa I. Iezzoni, M.D., M.S., *Using Severity Information For Quality Assessment*, Dec. 1989 QUAL. REV. BULL. 376; J. William Thomas & Marie L. F. Ashcraft, *Measuring Severity of Illness*, 26 INQUIRY 483 (1989).

[37] *Severity of Illness*, *supra* note 21, at 245; Smith, *supra* note 21, at 1116-17; *In-Hospital Mortality*, *supra* note 36.

[38] Mark S. Blumberg, M.D., *Biased Estimates of Expected Acute Myocardial Infarction Mortality using MedisGroups Admission Severity Groups*, 265 JAMA 2965 (1991); Smith, *supra* note 21, at 1118-19.

is R^2, a statistic that summarizes the degree of "fit" of a regression model. R^2 ranges from 0% to 100%. Green and his colleagues reported that R^2 values for a simulated HCFA-type model averaged 2.5% across five diagnostic categories.[39] The same study found that adding an explicit measurement of severity, the Severity of Illness Index, increased the R^2 to 21.5%.[40] Jennifer Daley and her colleagues compared models incorporating two such instruments—the Medicare Mortality Predictor System and APACHE II—and reported R^2 values for stroke patients of 25.3% and 22.0% respectively.[41] Lisa I. Iezzoni has also reported good predictive capacity with a model using MedisGroups.[42]

While these preliminary comparisons cannot be considered benchmarks, since they came from different studies and used different criteria for sampling patients, certain patterns emerge. It is clear that the R^2 statistics for the simulated HCFA model, derived exclusively from hospital claims data, are quite small. By contrast, predictions based on any of the well known severity of illness systems are, in general, about an order of magnitude larger. Additionally, even the severity adjusted measures have R^2 statistics that are generally in the 15%-25% range, implying that most of the variation in mortality rates remain unexplained even by the best available measures.

B. *Data Accuracy*

When interpreting the fact that a particular providers' mortality rate is much higher than predicted, the possibility that the result may be due to data errors should always be considered. Experienced data analysts know that when an "outlier" appears in a set of data, the first thing to check for is a possible recording error.[43] Some of the health care databases that are used to generate published mortality statistics contain many inaccura-

---

[39] *Severity of Illness*, *supra* note 21, at 243.

[40] Susan D. Horn, Ph.D. et al., *Measuring Severity of Illness*, 21 MED. CARE 14 (1983).

[41] Jennifer Daley, M.D. et al., *Predicting Hospital-Associated Mortality for Medicare Patients*, 260 JAMA 3617 (1988).

[42] Lisa I. Iezzoni, M.D., M.S. et al., The Ability of MedisGroups and Clinical Variables to Predict Cost and In-Hospital Deaths at 143-46 (July 1, 1988) (research report, on file with the author).

[43] F.J. Anscombe, *Rejection of Outliers*, 2 TECHNOMETRICS 123 (1960).

cies. The literature reports error rates for patients' principal diagnoses ranging from 18.5%[44] to 42.8%,[45] and about 26% for secondary diagnoses (comorbidities).[46] Doubts have been expressed about the coding accuracy of two other data elements used as mortality risk adjusters: "transfer source" and "type of admission."[47] Moreover, research indicates that such error rates are not uniform across hospitals.[48]

C. *Which Results are Right?*

Different clinical studies comparing treatments sometimes reach opposite conclusions because of subtle methodological differences. For example, several published studies using claims databases indicated that mortality rates following transurethral resection of the prostate (TURP) exceeded the rates following a different procedure, open prostatectomy.[49] These results surprised urologists since TURP was considered the safer procedure than open prostatectomy. Some urologists speculated that the findings could be due to the fact that TURP is more frequently performed on high-risk patients than open prostatectomy.[50] John Concato and his colleagues tested this hypothesis by examining medical records of 252 men who underwent either operation and found that, after controlling for the full extent of patients' illnesses (comorbidities), the mortality rates for the two procedures were indistinguishable.[51]

Similarly, varying approaches to studying provider outcomes often yield different results. Several studies have found

---

[44] Richard F. Corn, *The Sensitivity of Prospective Hospital Reimbursement to Errors in Patient Data*, 18 INQUIRY 351, 354 (1981).

[45] Linda K. Demlo, Ph.D. et al., *Reliability of Information Abstracted from Patients' Medical Records*, 16 MED. CARE 995, 998 (1978).

[46] *Id.* at 999; Allan N. Johnson & Gary L. Appel, *DRGs and Hospital Case Records*, 21 INQUIRY 128, 133 (1984).

[47] Elizabeth Gardner, *UB-82 Forms Offer Wealth of Information, Misinformation*, MOD. HEALTHCARE, Sept. 24, 1990, at 18, 24.

[48] Susan S. Lloyd, R.R.A. & J. Peter Rissing, M.D., *Physician and Coding Errors in Patient Records*, 254 JAMA 1330, 1333 (1985); Naomi S. Soderstrom, *Are Reporting Errors Under PPS Random or Systematic?*, 27 INQUIRY 234 (1990).

[49] Noralon P. Roos & Ernest W. Ramsey, *A Population-Based Study of Prostatectomy*, 137 J. UROLOGY 1184, 1187 (1987); John E. Wennberg et al., *Use of Claims Data Systems to Evaluate Health Care Outcomes*, 257 JAMA 933, 935 (1987).

[50] John Concato, M.D., M.S., M.P.H. et al., *Problems of Comorbidity in Mortality after Prostatectomy*, 267 JAMA 1077, 1080-81 (1992).

[51] *Id.*

that when lists of hospital mortality outliers generated from databases are examined further using severity of illness adjustments, the lists change. In some studies many of the "outliers" flagged by a method like that of HCFA ceased to be outliers when severity of illness was incorporated.[52] Edward L. Hannan, the developer of the Cardiac Surgery Reporting System ("CSRS") used in New York State, found that claims data alone generated a different set of cardiac surgery mortality outlier hospitals than a more detailed database that included clinical risk factors.[53]

Even when the same database is used, various ways of specifying the statistical model can yield markedly different lists of outliers. For example, Green and his colleagues compared the lists of outlier hospitals generated for hospital admissions in 1986 in two of HCFA's annual mortality reports. The comparison revealed that HCFA's methodological changes between the 1987 and 1989 mortality reports gave rise to new assessments of the outlier status of individual hospitals, even when the same hospitalizations were studied. Of the 126 hospitals designated as high mortality outliers in 1986, according to HCFA's 1987 report, 53, or 42%, were no longer considered to have been outliers for the year 1986 in the 1989 report.[54] Such modification of important study results as methods change may either indicate progress or suggest that the findings are too sensitive to methodological details.

VI. New York State's Cardiac Surgery Mortality Report

An ongoing study by the New York State Department of Health is evaluating inpatient hospital mortality rates following open heart surgery.[55] The project relies on a database, CSRS, containing data on patients' risk factors for post-surgical mortality and their in-hospital survival outcomes. Among the databases currently used to compare providers, the CSRS study probably represents the best design yet implemented. Therefore,

---

[52] Robert W. DuBois, M.D., Ph.D. et al., *Hospital Inpatient Mortality*, 317 New Eng. J. Med. 1674, 1676, 1678-79 (1987); *Hospital Mortality*, *supra* note 21; *Severity of Illness*, *supra* note 21, at 245; Smith, *supra* note 21, at 1116-19.

[53] Hannan, *supra* note 7, at 2773.

[54] Green, *supra* note 33, at 11.

[55] Hannan, *supra* note 7, at 2773.

as the leading example of this kind of research, the New York State study merits careful evaluation.

The development and some applications of the CSRS have been described elsewhere.[56] Briefly, the database consists of records of all adult open heart surgical discharges from New York State hospitals since 1989. The form used to collect the information, developed by a panel of surgeons and statisticians, includes information on demographics, types of procedures and clinical risk factors.

In comparison with HCFA's study, the New York cardiac surgery investigation has some distinct advantages. First, while HCFA has had to rely on a database designed for hospital billing rather than risk assessment, the CSRS was designed specifically for research on cardiac surgical mortality. Second, the New York State study focuses on a more homogeneous patient population—those undergoing cardiac surgery—than HCFA's study of all Medicare patients hospitalized in the United States. Third, hospital care involving an intensive therapeutic intervention like open heart surgery would almost never occur during admissions for terminal, hospice-type care, thus avoiding a major potential confounding factor in HCFA's study. Finally, the severity of illness measure used in the CSRS is based on a broad array of demographic and clinical factors associated with post-surgical survival and this measure is supported by considerable scientific evidence.

On the other hand, HCFA's study had the advantage of access to dates of all Medicare beneficiary deaths independent of the inpatient data and could therefore hold survival durations constant for all patients at 30, 90 and 180 days. The New York State database is strictly an inpatient record and is limited to studying deaths that occurred in the hospital only.

There are also potential problems in the design of the CSRS study. The group of cardiac patients studied in the CSRS is not entirely homogeneous. For example, patients undergoing emergency open heart surgery for "disasters," such as gunshot or knife wounds, with a mortality rate of 43%, are included with those having only bypass surgery, with a mortality rate of 4%. It

---

[56] *Id.*; Edward L. Hannan, Ph.D. et al., *Coronary Artery Bypass Surgery*, 29 MED. CARE 1094 (1991); New York Department of Health, *Cardiac Surgery Research Results Encourage Improvements in Performance*, EPIDEMIOLOGY NOTES Feb. 1991.

is unlikely that a single statistical model can accommodate scheduled bypass surgery to relieve coronary artery disease and cases of trauma from gunshots. Instead, "disasters" should probably be studied separately.

In addition, while the list of risk factors included in the study conforms to those used in other research on cardiac surgery, it was not exhaustive.[57] According to the developers, certain risk factors, such as ventricular dysfunction and severity of coronary arteriosclerosis, that are known to increase the risk of post-surgical mortality were not incorporated into the CSRS because their inclusion would have been too difficult or expensive.[58] The first published study based upon the CSRS documented the construction of the databases, an evaluation of risk factors and an estimation of a model to predict mortality.[59] No hospitals were identified in the article. However, a press release issued by the New York State Department of Health coinciding with the study's publication did identify the hospitals by name, ranking them from one to thirty according to their risk-adjusted mortality rates. This ranked list of New York State's surgical centers appeared in the *New York Times* on December 5, 1990.[60]

While consumer groups applauded and hospital officials expressed concern about the release of the mortality rate rankings of hospitals, issues regarding the scientific validity of the rankings received little coverage. On the one hand, the scientific report on the CSRS showed that a hospital's predicted mortality rate could be precisely determined from the CSRS using a 95% confidence interval, only to within plus or minus 2-3%.[61] On the other hand, the hospital rankings released by New York State and published in a chart by the *New York Times* differed by as little as .01%, a difference that is more than 100 times smaller than the threshold of statistical significance.[62]

---

[57] Victor Parsonnet, M.D., F.A.C.C. et al., *A Method of Uniform Stratification of Risk for Evaluating the Results of Surgery in Acquired Adult Heart Disease*, 79 CIRCULATION I-3 (1989).

[58] Hannan, *supra* note 7, at 2772.

[59] *Id.*

[60] Lawrence K. Altman, *Heart-Surgery Death Rates Decline in New York*, N.Y. TIMES, Dec. 5, 1990, at B10.

[61] Hannan, *supra* note 7, at 2772.

[62] Altman, *supra* note 60.

VII. A New Precedent: Naming the Physicians

On December 17, 1990 a *New York Newsday* reporter filed a request with the New York State Department of Health under New York's Freedom of Information Law for physician-specific mortality data derived from the CSRS. The Health Department denied the request initially and on administrative appeal. The basis for denial was that disclosure of the physician names could constitute a violation of the Personal Privacy Protection Law.[63] *Newsday* appealed to the New York State Supreme Court. An affidavit from the Health Department opposing the appeal again argued that release would invade physicians' privacy and that the invasion would be unwarranted "because of the potential that the data would be misunderstood and misused by the public, resulting in significant adverse impact upon the physicians identified by the data, with little public benefit."[64]

While arguing that the public would misunderstand the data, the Health Department simultaneously represented that the data themselves were accurate. By representing the validity of the data, while questioning the ability of consumers to understand them, the Department left itself open to the following retort by the court: "In other words, the State must protect its citizens from their intellectual shortcomings by keeping from them information beyond their ability to comprehend."[65] In finding for *Newsday*, the court concluded that physicians had little reason to expect that their surgical results would be kept private and added that "even if there was a legitimate privacy expectation, the interest of the public outweighs it."[66]

Because the Health Department raised no question about the validity of its data, the deliberations proceeded without discussion of the limitations of the physician comparisons. However, these limitations were apparently quite serious. The following is an excerpt from a statement issued subsequently by the Department:

---

[63] Letter from Peter Slocum, Records Appeals Officer, N.Y. State Dep't of Health, to Nancy E. Richman, Senior Staff Counsel, Newsday (Feb. 15, 1991) (on file with the author).

[64] Respondent's Affidavit in Opposition at 3, Newsday v. N.Y. State Dep't of Health, 19 Media L. Rep. (BNA) 1477 (N.Y. Sup. Ct. 1991) (No. 3406-91).

[65] *Newsday*, 19 Media L. Rep. (BNA) at 1478.

[66] *Id.* at 1479.

> To examine the stability of the data, the risk-adjusted patient mortality rates of cardiac surgeons in 1989 and 1990 were compared to determine how valuable 1989 information was in predicting 1990 performance. The results of those comparisons demonstrate that the 1989 risk-adjusted mortality rates cannot be used to predict 1990 performance at a surgeon level. For example, of the 29 surgeons with risk-adjusted mortality rates in the highest quartile in 1989, 9 (31%) had risk-adjusted mortality rates that were in the lowest half in 1990, and six (21%) had risk-adjusted rates in the lowest quartile in 1990. This analysis is based on the 118 cardiac surgeons performing coronary artery bypass graft operations in New York State in both 1989 and 1990.
>
> Of the 30 surgeons with risk-adjusted rates in the lowest quartile in 1989, twelve (40%) had risk-adjusted rates in the upper half in 1990, and three (10%) had risk-adjusted rates in the highest quartile.
>
> Overall, the correlation coefficient (Pearson) for the surgeon-specific rates in 1989 and 1990 was a very low .11. This means that using a surgeon's 1989 risk-adjusted rate to predict the 1990 rate was only slightly better than pure chance. The correlation coefficient for only those surgeons with annual case volumes of 50 or more improves to a modest .27, still too low to be confident in the predictive power of the model applied to surgeon performance.[67]

One can only speculate as to whether the *Newsday* court's balancing of the public's interest against the privacy rights of physicians would have turned out differently if the court had considered the serious flaws in the data. If the public's interest in the data stems from their usefulness in evaluating providers, then the serious limitations of the data's validity would certainly diminish this interest and should have been taken into account.

VIII. Informing The Court: The Need for Evaluation of Outcome Statistics

Where provider-specific outcome data could become part of litigation in either malpractice or informed consent cases, it is important that courts be fully aware of the limitations of these data.[68] Otherwise, the mere fact that the data have been disseminated may lead courts to conclude that physicians have a duty to disclose them to patients. The duty to disclose information about risks associated with medical interventions cannot be in-

---

[67] Inter-office Memorandum from Peter Slocum, Dir. of Publ. Affairs, N.Y. State Dep't of Health, to Patricia Montone 4 (Dec. 12, 1991) (on file with author).

[68] William J. Curran, *The Acceptance of Scientific Evidence in the Courts*, 309 New Eng. J. Med. 713 (1983).

dependent of the scientific basis on which that information rests. Where scientific evidence is strong, the duty to disclose a risk must be greater than in instances where there is little credible basis for a risk.

At a minimum, users of the data should have access to information on the validity and reliability of provider comparisons. The "goodness-of-fit" of the mathematical models used to control for patient risk factors should be disclosed along with the publication of the mortality rates. As noted above, mortality prediction systems are often assessed in terms of R2 statistics that allow for quantitative comparisons.[69] Another useful statistic for evaluating the model would be the predicted mortality rate for patients who died. If the model fits the data well, most predicted deaths should come from this group.

Finally, when deciding questions related to disclosure of such data—whether public disclosure under freedom of information laws or disclosure by physicians to their patients—courts should be briefed on the intrinsic limitations of outcome comparisons derived from observational databases. In particular, doubts must exist when hospitals and surgeons are compared using the experience of patients who are not randomized. Many factors that influence patients to select a particular surgeon create enormous potential for misleading outcome comparisons across surgeons, such as cardiologists' referrals and surgeons' reputations. Statistical adjustment alone cannot overcome the fact that hospitals or surgeons who take on tough cases tend to look bad in a mortality report.

To decide whether the release of comparative outcome statistics that are only partially adjusted for severity of illness is in the public interest, courts should consider the potential for unwanted repercussions with the publication of such rankings. For example, a recent *Newsday* article reported that "[s]everely ill heart patients are finding it increasingly difficult to get surgery, many doctors say, because some surgeons and hospitals are refusing to take patients they fear will lower their standing in state mortality statistics."[70] Interviews with cardiologists and cardiac surgeons revealed that physicians in New York State "do not believe the system gives enough credit for operations on the

---

[69] *See* Hannan, *supra* note 7.

[70] Zinman, *supra* note 12.

sickest patients."[71]

The perception that the severity of illness adjustment is inadequate may have been reinforced when the hospital-specific mortality rates were released. Many in the medical profession noticed that the top ranked hospital in 1990 had been number 24 of 30 in 1989. One explanation offered was this hospital's decision "not to do as many tougher cases" after seeing the 1989 data.[72] This explanation casts further doubt on the study's validity; if the data were adequately adjusted for patients' severity of illness, then selecting easier cases would not improve a hospital's rank.

C.J. McDonald and Sin L. Hui have wisely observed that:

> Regardless of how they are developed, policy makers should be slow to use 'performance' measures to reward or punish for two reasons. The first is Byar's argument about using databases to compare treatments. The hospital and physician are in some sense a treatment, and the selection of hospital/physician is confounded by the patient's condition—especially when we consider the patient referral process. The second is a principle from Deming's continuous quality improvement plan. Measures of performance should be used only to improve the *system*. All such measures have defects. If used for reward or punishment, they divert worker energy from improving the system to criticizing, or gaming, the performance measures.[73]

CONCLUSION

Any trend toward the introduction of provider-specific outcome statistics as evidence in informed consent or malpractice litigation can only be expected to increase defensive behavior on the part of physicians and hospitals. While research into variations in the effectiveness of care should continue, premature utilization of the early products of such research has considerable potential to misinform consumers and demoralize health care providers.

---

[71] *Id.*

[72] Altman, *supra* note 60.

[73] Clement J. McDonald & Sin L. Hui, *The Analysis of Humongous Databases*, 10 STAT. IN MED. 511, 514 (1991) (emphasis in original).