Following the Man on the Clapham Omnibus: Social Science Evidence in Malpractice Litigation

Richard Lempert
University of Michigan Law School, rlempert@umich.edu

FOLLOWING THE MAN ON THE CLAPHAM OMNIBUS: SOCIAL SCIENCE EVIDENCE IN MALPRACTICE LITIGATION

RICHARD LEMPERT

Following the Man on the Clapham Omnibus: Social Science Evidence in Malpractice Litigation

Richard Lempert[1]

Part I: *"That is not what I meant at all. This is not it, at all."* [2]

We are tourists in a London court, watching a trial. The plaintiff tripped and fell over a barrier the defendant had set up to protect wet cement. He claims that the barrier should have been higher, with blinking lights to be sure it would be seen. The defense is contributory negligence, and the plaintiff's case is not going well for him. The defendant has established through two witnesses that when the accident occurred the plaintiff was talking excitedly into his cell phone and not looking where he was going. At the moment plaintiff's attorney, a Queen's Counsel, is arguing about the admissibility of testimony he is offering to show that his client's inattention was not negligence. We pick up the case here:

QC: Detective Jones, can you tell the jury what you did on the afternoon of January 9, a year to the day after my client's horrible accident?

WIT: Yes, at your instructions I got on the omnibus for Clapham. The bus was empty except for one man.

QC: Can you tell us about that man?

WIT: Yes, his name, as I learned when I spoke to him, was Mr. Xbar. He was of average build, and average looks, and to judge by his conversation of average intelligence. He was dressed in typical London style and carrying an ordinary

---

[1] Eric Stein Distinguished University Professor of Law and Sociology. I wish to thank Richard Friedman and Lisa Kahn for their comments on an earlier version of this paper.

[2] T.S. Elliot, The Love Song of J. Alfred Prufrock

black umbrella.

QC: What happened next?

WIT: The bus stopped and he got out, so I followed him.

QC: What did you see?

WIT: He hadn't gone more than twenty feet when he took out his cell phone and began an animated conversation. He wasn't looking where he was going. Indeed, there was a barrier on the sidewalk just like the one in this case, and he certainly would have tripped over it had I not yelled out.

QC: Thank you very much. Your Honor, I think the plaintiff is entitled to a directed verdict on the contributory negligence issue, for I can cite you hundreds of cases in which the high court has told us that negligence law expects no more care than that which can be expected of "the man on the Clapham omnibus." Now, for the first time we have shown that man's actual behavior - we have shown empirically that talking excitedly on a cell phone and not paying attention to the sidewalk is ordinary behavior.

Can there be any doubt what happened next in this case—after the judge stopped laughing? Surely the directed verdict was denied, and almost as surely upon the defendant's motion, the detective's testimony was stricken.

It may seem easy to separate fact from law in negligence cases or, what is the same thing, proof from norm. If science and technology allow us to be more certain that a legal standard has been met, the urge to embrace what science offers is strong—curbed only by questions about the adequacy of the science and the discipline of cost-benefit analysis. Using science to ascertain whether the elements of a claim or charge are met is most often not only justified but desirable. In a criminal case, for example, the defendant's involvement in a crime must be proved beyond a reasonable doubt. When DNA from the criminal is available, as it often is in rape cases, it is the gold standard for proof of identification,

replacing the prior best evidence—confident eye witness identification.[3] It is sufficient to secure convictions with no other proof of identity, and it can acquit those once convicted, even when eyewitnesses are adamant that the original guilty verdict was correct. One reason DNA evidence has this power is because the norm requiring a defendant's criminal involvement to be proved beyond a reasonable doubt expresses a value that seems unlinked to modes of proof. As a matter of social policy, we want to be able to feel sure that the defendant "did it" no matter how easy or difficult it is to show that he did.[4]

But other legal norms, I would argue, co-evolved with the technology available to prove them. With different technologies we would have different norms. This is transparent in the case of "the man on the Clapham omnibus;" his behavior was never meant to be proved by observation. Rather, this man is the judicially constructed image of a sane, sober but not extraordinarily gifted person who never takes unreasonable chances, and does nothing extraordinary, but does everything that is ordinary to perfection. Had it been possible to actually follow a sample of men alighting from the Clapham omnibus, English courts would have found a different image to convey what they meant by the reasonable man. Actual Clapham bus riders would on some occasions have been too careless and on others too cautious to do the work of tort law's reasonable man.

The "customary physician behavior" standard courts use to assess whether a doctor's actions are malpractice has more in common with the Clapham omnibus standard than it does with the requirement that a criminal's identity be proved beyond a reasonable doubt. The customary practice norm evolved, I suggest, along with the methods available for proving not just customary behavior but also reasonable medical practice. Had either been different, I expect different norms would have evolved. Hence when I read proposals to identify customary physician behavior through social science,

---

[3] Occasionally fingerprint evidence was available, but this was relatively rare, and the power of blood markers was largely confined to exclusions.

[4] The norm may, however, not be independent of modes of fact finding. If juries in fact never convicted when there was objectively some reasonable doubt, e.g. some small but non-negligible objective probability that one eyewitness was mistaken—a more lenient standard of proof in criminal cases might have arisen.

with the implication that what empirical investigation may show it is not what people think (or want to believe) it is, my instincts say "proceed with caution."

I might feel differently if legal evolution were immediate and responsive to social needs. Then, if empirical investigation promised to yield true facts, I could see substantial virtue in confronting the legal system with them and trusting it to readjust. This can happen. The shift in malpractice law in many jurisdictions from localized to regional or national standards of customary physician behavior is an example of malpractice law responding to greater factual certainty. As standards for diagnosing and treating particular maladies became more precisely defined and nationally disseminated, many courts and legislatures thought it unreasonable to allow injuries to go uncompensated because few, if any, local physicians followed best practices. In particular, when doctors held themselves out to be specialists, they began to be treated as part of a national community of specialists rather than a local community of physicians. But legal evolution of this sort is ordinarily a slow process, and it is likely to be made slower by physician lobbying and a legal culture which assumes that the law should not demand more of doctors than what good practice, as evidenced by what most doctors do, entails.

The assumption that good practice is what most doctors do is, however, tenuous when medical science advances so rapidly that physicians have trouble keeping up, even without natural human tendencies to follow familiar protocols and practice structures that can pose barriers to the quick adoption of best medical practices. Caught between what best practice demands of doctors and what doctors can reasonably be expected to do, the common law has arrived at one of its disingenuous but highly practical compromises. Malpractice is behavior that falls short of what the average or typical physician in a relevant community would do, but the typical physician, like the man on the Clapham omnibus, is a construct, not a person. Average or typical behavior has no practical meaning apart from the ways it can be shown. In malpractice cases, the behavior expected from the typical practitioner is usually shown by expert testimony. Hence the law's typical physician often behaves more competently than the average community physician. This is because plaintiff's experts testify that behavior which may be more competent than what is usually achieved is average. They do this by confusing the medical profession's aspirations for treatment with behavior that always lags aspirations, and asserting a

community standard that is informed by aspiration as well as behavior.

Juries are no dummies. They know that aspirations are not always achieved, and they know that plaintiffs have incentives to exaggerate the typicality of best practices. At the same time they can assess knowledge of best practice norms, the feasibility of adhering to them and the costs of persisting in next best or even far from best practices. The result I expect—although I cannot show this empirically—is that verdicts reflect information both about the standards physicians actually attain (and in some cases not even that) and the standards they should attain, as jurors weigh conflicting expert testimony on existing standards of behavior. It is likely that the greater the apparent health care advantage of aspirational over behavioral norms and the easier they are to attain, the more likely fact finders are to treat aspirational norms as behavioral ones. If this is what occurs, I see it as often a good thing, which helps explain why norms which privilege typical physician behavior in malpractice cases have been able to endure.

Importantly, whether the law holds physicians to aspirational or behavioral norms, the legal system is still delegating the task of establishing normative standards to the medical profession. Moreover, the norms I have called aspirational are not aspirational in the sense that they are the practice standards that leading physicians hope will someday be achieved; rather they are practice standards that experts, presumably testifying in good faith, believe are now being achieved by ordinarily competent physicians. At least this is the lesson I draw from Meadow's and Sunstein's interesting work.[5] Meadow and Sunstein found that when two groups of experts were asked to estimate the time lapse between the arrival at emergency rooms of children presenting with bacterial meningitis and the administration of antibiotics, their median estimates were 46 minutes and 80 minutes, although an empirical investigation of two emergency rooms indicates that the actual average lapse is 120 minutes.[6] Meadow and Sunstein argue that juries would do a better job if they were regularly provided with data

---

[5] W. Meadow and C.R. Sunstein, *Statistics, Not Experts,* 51 Duke L. J. 629 (2001); see also, W. Meadow, *Operationalizing the Standard of Medical Care: Uses and Limitations of Epidemiology to Guide Expert Testimony in Medical Negligence Allegations,* ___ Wake Forest L. Rev. ____.

[6] Meadow and Sunstein, supra note 5, at 638.

like that which they report, but Meadow and Sunstein reach this conclusion only by bracketing the question of whether the law should incorporate or sometimes improve on, the existing standards of care.[7] I don't think this question can be so casually put aside, for the wisdom of their preference for statistical over expert knowledge turns in large measure on this issue.

My preferences are for a system which moves actual practice toward standards that experts feel are not just reachable but ordinarily reached. Recall, we are only talking about situations in which the gap between the aspirational and behavioral norms is likely to have resulted in serious harm to a patient. Unless the harm were serious a malpractice case would not be brought, and unless the gap appears more likely than not to have caused the harm, showing the gap will not justify a plaintiff's recovery.[8] Meadow and Sunstein might object that my preferences are fine, but they are not the law's preferences, and that they take the legal standard seriously and are seeking to provide the law with the evidence the law seeks. The position is not unreasonable but brings me back to the proposition with which this paper begins: the legal standard co-evolved with available means of proof, and a somewhat inflated expert view of what ordinary physicians do may be more consistent with what the law's community practice standard has meant historically than is empirical research, even if the latter portrays more accurately actual physician behavior.

Part II: *There are three kinds of lies: lies, damned lies, and statistics.*[9]

We live in an age where literalism is a chic form of legal interpretation. Taking malpractice

---

[7] *Id.* at note 4.

[8] It might, however, contribute to one, as a jury might think a defendant's actions are more aberrant than they are, and this might affect their judgment of other kinds of alleged negligence.

[9] The New International Dictionary of Quotations, Third Edition, (selected and annotated by Margaret Miner and Hugh Rawson, Signet, 2000, p.434); quote was attributed to Benjamin Disraeli by Mark Twain in M. Twain Autobiography (1924).

law's deference to community standards literally, as courts are likely to take it, there is nothing radical about proposals to inform malpractice litigation with empirical data. For the literalist court or academic commentator, problems center on the feasibility of empirical investigation and the reliability and utility of empirical data. Here too I have my doubts. There is no better way to explain them than by commenting on the examples and suggestions that the symposium papers offer.

The best way to gather information about how physicians actually behave is, in principle, to observe them or to analyze data from those who have. This is what Meadow and Sunstein propose in their Duke paper and what Meadow proposes in his contribution to this symposium. Although Meadow and Sunstein give an interesting example of how this might be done by reviewing case records and extracting from them data showing that the judgments of experienced experts may be far different from what the coding and counting of case histories reveals, studies based on observations of physician practice are fraught with practical and methodological difficulties. First, there is the expense of doing the needed studies. Each malady that might be the subject of a malpractice suit would need its own survey. Where local standards of competence are the touchstone for determining malpractice, only studies of local practice will bear directly on the crucial legal issue. Rarely will such studies be available, and even when they are, the number of cases observed is likely to be too small to refute expert judgments about the standards that are ordinarily achieved by competent local doctors. For example, Meadow and Sunstein's Chicago emergency room data show that the time lapse from patient presentation to the start of antibiotic therapy is considerably longer than what the experts they queried believed, but in an earlier version of their paper[10], the average expert judgment was well within the 95% confidence interval of the average time in their study. Only by adding perhaps non-comparable cases from other states were they able to conclude at conventional levels of statistical significance that their experts had on average underestimated the typical time to treatment. In the version of their paper published in the *Duke Law Journal*, the differences are significant, but a different statistical test is

[10] Meadow and Sustein, *Statistics Not Experts,* John M Olin Law & Economics Working Paper, No. 109.

used.[11]

Even if the differences they report were not significant, it would not mean that the data they collected would be irrelevant in a malpractice suit against a Chicago area doctor. Knowing the range of actual response times tells a jury something about actual practice in the locale, even if it does not justify substantial confidence that the expert testimony misses the mark. There is, however, a further problem in treating the empirical and expert evidence as inconsistent. The cases considered by the experts and those counted by observers may not be the same. Experts asked to estimate the average time from arrival in the ER to treatment for children presenting with bacterial meningitis (or any other malady) may be thinking of cases with a certain array of symptoms. The observed cases reviewed which were presumably selected because bacterial meningitis was diagnosed and antibiotic treatment begun, may have arrived at the ER displaying a wide range of symptoms, some of which may have been typical of bacterial meningitis and some of which may have been more suggestive of other illnesses. The expert judgments of average time to treatment for typical cases may be accurate, whatever a *post hoc* review of time to treatment for all cases diagnosed with meningitis indicates. If the patient whose treatment led to the malpractice suit presented with typical symptoms, something an expert might testify to, empirical data based on cases with a range of presenting symptoms might present a misleading picture of local practice standards.

In theory, case review studies can account for different case characteristics by coding more variables, such as indicators of symptoms upon presentment. Classifying cases by presentment symptoms (or any other relevant variable) would, however, further reduce the number of relevant cases and give yet less reliable estimates of average time lags. More importantly, coding more variables increases occasions for subjective judgements, already a weakness of case review evaluations.

Subjectivity problems begin with the case record. All that can be evaluated in a *post hoc* case review is what is written in the medical record, and what is written will often reflect a myriad of subjective judgments. Even a variable that seems as objective and easy to define as time to treatment might embody considerable subjective judgment. If patients are asked when they arrived in the ER,

---

[11] See the text accompanying notes 23 - 26 *infra*.

they may overestimate the time they have been waiting. If medical charts are filled in after treatment, the estimated time of antibiotic administration may deviate from when the drugs were actually administered, or be affected by cognitive biases. For example, Meadow and Sunstein report that the actual median time to treatment in the 93 Chicago area cases for which they had data was precisely two hours. I wonder if this round number is a coincidence or if a number of the records Meadow reviewed gave two hours as the time to treatment, reflecting not actual times to treatment but rounded estimates. Subjectivity is also likely when coding record data. If only trained physicians can do the coding, the costs of coding (including dual coding to ensure reliability) will be considerable, with out of pocket financial costs compounded by the costs of diverting physicians from patient care. Additional problems exist if either ER record makers or coders know that what they find will be admissible in malpractice suits. Temptations to shade entries to protect physicians are obvious.

If national practice standards are at issue, the difficulty of acquiring enough cases to get reliable estimates of typical treatments is diminished, but other problems emerge. To estimate empirically practice standards across the nation, not only should practice settings be sampled randomly,[12] but the information acquired in the different settings must be comparable. If different information was collected in different locales or if local coding practices differed (e.g. some E.R.'s coded presentment time as the time the patient walked through the door, others as the time the patient registered with intake and others as the time the patient first saw a medical screener, it might be impossible to reconcile data from different locations.)

Of course, a national sample might miss the mark entirely, for when courts speak of applying national practice standards, they do not seem to mean that the conduct at issue should be judged against a national mean performance measure, in which the decisions of a family physician in West Virginia count as much as those of a specialist in New York City. National standards do not mean average standards across the nation but rather refer to nationally accepted standards of medical competence. This confronts empirical investigators with additional problems. If national standards do

[12] This does not necessarily require a simple random sample; some type of stratified probability sample could be used.

not reflect the typical performance of all physicians, which subgroups of physicians should one study? Since it is specialists who most commonly are held to national standards, one might think the target group for study should be all specialists in the malady in question. We must still, however, confront the question of whether all specialists should count equally in determining acceptable performance, and if coders are working from records there is the additional problem of determining whether a treating physician was a specialist. Perhaps the best we could do would be to study a small number of admittedly exemplary medical care providers and treat the quality of their medical care as an upper bound on what competent treatment entails. Any physician who acts as they would, surely must be attaining the minimum national standards of competence. The problem is that the hard cases will involve physicians who weren't quite as savvy as the group of exemplars.

Meadow and Sunstein's proposal is further limited because not all medical treatment decision making is as visible as the treatment of children presenting to E.R.s with bacterial meningitis, and not all treatment information is as easy to code as the time lapse between presentment and treatment. Gathering reliable empirical information will be especially difficult for office treatments and diagnosis, given the likely reluctance of physicians to make their office records available and inconsistencies in how different offices record similar information.

Meadow and Sunstein say their proposal "rests partly on the claim that the use of statistical data will greatly simplify litigation and reduce the role of strategic behavior in the litigation process."[13] If statistics were not socially constructed, and if we could provide juries with indisputably valid statistics, then statistical evidence could simplify litigation and perhaps reduce the role of strategic behavior. Collecting good (which is to say valid and reliable) statistics is, however, a difficult and highly contestable process, especially where the actors who generate the statistics have a stake in what the evidence portrays. Rather than simplifying trials, statistical evidence of physician behavior will introduce complicated new issues into malpractice trials and create battles of experts that are at least as difficult to resolve as the battles that now occur between "standard of practice" experts.

The Meadow and Sunstein proposal is a particularly ambitious proposal for introducing

---

[13] Meadow and Sunstein, supra note 4 at note 4.

empirical evidence of standards of practice studies into malpractice litigation. The data they would have us acquire is better geared toward establishing actual practice standards than the data the other symposium authors would collect, analyze and admit. The very ambition of the Meadow and Sunstein proposal makes it easy to criticize. The other proposals are more feasible because their data collection requirements are less onerous. Nevertheless, they share problems with the Meadow and Sunstein proposal, and have some of their own.

Hartz *et al's* proposed use of survey data[14] makes sense if the data are used as they suggest—to aid settlement. Mail surveys can be done quickly and cheaply, and their results can serve as a reality check for the parties, particularly if responses are overwhelmingly in one direction. But mail surveys are not a good way to generate admissible practice standard evidence. All surveys are open to biases induced by question wording and question order.[15] When the parties are using survey results only to inform their negotiations, consensus on question construction is likely to be easier to achieve than when the data may be offered to a jury. Parties to a settlement survey will want to make the survey fair since if it is obviously unfair, disappointing results will not induce the disappointed party to settle. But if survey results will be given to a jury, competing surveys might be offered or, if a court pressured the parties to develop a common instrument, substantial wrangling would occur over how to word and order questions.

Moreover, physicians responding to a settlement survey would expect their honest answers to facilitate settlement. But if they knew their responses would figure in a malpractice trial, professional solidarity might color some replies. Even a small proportion of strategic responders could substantially distort measures of average performance.

---

[14] A. Hartz, J. Lucas, T. Cramm, M. Green, S. Bentler, J. Ely and S. Wolfe, "Physician Surveys to Assess Customary Care in Medical Malpractice Cases." My example is from an earlier version of the paper presented at the symposium. The current paper is T. Cramm, A. Hartz and M. Green, "Ascertaining Customary Care in Malpractice Cases: Asking Those Who Know." _____ Wake Forest L. Rev. _____.

[15] H. Schuman and S. Presser, 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context.* New York: Academic Press.

Mail surveys have other weaknesses, many of which Hartz and his coauthors recognize. Not only are response rates often low, but they may also be biased —busier physicians, for example, might be less likely to respond and those most hostile to malpractice plaintiffs might be more likely. Also the survey instrument is unlikely to capture all relevant aspects of the case in question and, as in the survey Hartz reports, results might be equivocal, with different physicians reaching different conclusions. Indeed, in any case close enough to go to trial, mixed responses are almost guaranteed. The legal implication of mixed responses may not be clear. If a substantial minority of respondents would have acted as the defendant did, the defendant can argue that competent professionals have not settled on one best treatment, and there is no medical consensus. Because defendants can turn minority approval to their advantage, plaintiffs are unlikely to cooperate in surveys designed to provide trial evidence, but will instead seek to attack them as biased and unreliable. Except where a survey showed an overwhelming consensus, there is little reason to believe that confusion would diminish or more just outcomes would result. Where surveys did show an overwhelming consensus, settlement would be likely, which is how Hartz *et al* see surveys as being most effectively used. I concur in this judgment.[16]

Surveys can also be useful when they show dissensus. In addition to his study of time to treatment in meningitis cases, Meadow, in his paper, refers to two survey studies he participated in as examples of empirical research that might be used to establish standards of care in malpractice cases. One finds that despite the existence of practice guidelines that recommend the use of Ribavirin in children with respiratory syncytial disease most pediatric critical care physicians would not use them,[17] and the other study argues that physicians cannot fall short of a standard of care by failing to recommend home apnea monitoring (HAM) for infants discharged from neonatal intensive care with a history of apnea because disagreement among physicians and nurses, both across and within neonatal

---

[16] To increase the pressure to settle the parties might agree that if their survey achieved a sufficiently high and unbiased response rate and if it revealed considerable consensus about the adequacy of a treatment, the results could be admitted into evidence if the case didn't settle.

[17] A. R. Zucker and W.L. Meadow, "Pediatric critical care physician's attitudes about guidelines for the use of ribavirin in critically ill children with respiratory syncytial virus pneumonia," Critical Care Medicine, Vol. 23, No. 4 (1995).

intensive care units, means there is no standard of care.[18]

As survey research goes, these studies are problematic. Samples are non-random, and in one study, respondents were invited to give the survey to their associates. In addition, the questions are quite general, and neither survey elicits opinions about whether there are special situations in which physicians who ordinarily don't employ the treatments required about would use them. In addition, the study of Ribavirin leaves one completely puzzled about how the American Academy of Pediatrics could recommend the use of Ribavirin in situations where most apparently expert specialists would not use them. The HAM study applies only to the sparse vignettes presented and has no obvious relevance to more factually rich clinical decisions, and by just counting noses, it fails to tell us anything about whether HAM use is favored by leading intensive care units that might be expected to embody the standards all clinics should follow.

Nevertheless, despite their limits both studies uncovered patterns that seem reliably probative of dissensus in the medical community. While I would not treat them as exonerating physicians who failed to use Ribavirin or didn't order HAM, they provide exonerative evidence that courts should take seriously. Hence, I support the use of such studies in the limited ways I suggest in Part III of this article. At the same time, even if the evidence of these studies is uncontradicted, I think malpractice plaintiffs should be able to prevail in the Ribavirin case if they can persuade a jury that the Academy of Pediatrics practice standards are sound and the profession is lagging unduly in adopting them, and in the HAM case if they can persuade the jury that in their case, significant harm could have been averted by HAM, and that the facts of their case, unlike the "plain vanilla" facts of vignettes, should have alerted the physicians to the need for HAM.

I also think a jury could reasonably conclude from Meadow's surveys that the medical profession was so divided on the efficacy of the treatments in question that it was not malpractice to refuse to order them even if harm ensued and experts testified that they believed good standard practice

---

[18] W. Meadow, D. Mendez, J. Lantos, R Hipps and M. Ostrowski, "What is the legal 'standard of medical care' when there is no standard medical care? A survey of the use of home apnea monitoring by neonatology fellowship training programs in the United States," Pediatrics 1992; 89: 1083-1088.

required the treatments. These studies have power to persuade because the surveys were distributed to a group of apparently expert respondents, response rates were high, the respondents had no reason to believe their responses would play a role in malpractice litigation, no fancy statistical analyses are needed to spot patterns and differences, and the results by any measure indicate substantial dissensus and even disapproval of the treatments in question. Conditions like these will not always exist, but when they do, surveys will have their greatest value as evidence of professional opinion.

Hall *et al* present a third perspective on gathering empirical evidence to inform medical malpractice decisions.[19] They propose using existing databases to shed light on customary practice standards, and they discuss a variety of possible data sources. This suggestion has some advantages over the proposals of Meadow and Sunstein and Hartz *et al*. Hall *et al* would use data collected for other purposes, so the costs of providing courts with information would be largely limited to analysis. Also a number of the data sets they reference are quite large, giving considerable statistical power to analyses based on their data and, in some measure, alleviating concerns over the lack of random sampling. To the extent there are no obvious biases in these samples, practices that are widespread among sample physicians are likely to be common among all physicians, even if one cannot be certain whether 40% or 60% of all doctors follow the practice. If the law can live with this kind of imprecision, findings based on very broad and apparently representative populations are likely to be a helpful guide to how large numbers of doctors are practicing.

But as Hall and his coauthors acknowledge, the use of the data sets they identify is not without problems. Using data collected without litigation in mind eliminates a potentially important source of bias, but it also has disadvantages. Adapting data compiled for another use to another purpose is commonly difficult, often frustrating and sometimes impossible. Moreover, practice variation databases, as Hall *et al* point out, usually focus on populations of patients rather than communities of physicians. These studies show too much patient variation and contain too few specifics to extract

---

[19] M. Hall, R. Anderson, R. Balkrishnan, D. Goff, S. Feldman, A. Fleischer, Jr., B. Mellen and W. Moran, "Measuring Medical Malpractice Patterns: Sources of Evidence from Health Services Research."

standards of care to apply in malpractice cases. Provider proficiency studies tend to be too institutionally specific to shed much light on community practice standards and often focus more on cost issues than on competence. Medicaid databases have serious biases, as they collect data only for indigent patients. Medicare databases are also biased, as they are confined to the elderly, and they contain little prescription drug information. None of these data bases is designed to make it easy to extract information for use in malpractice litigation.

Potentially the most useful of the databases referenced by Hall *et al* is the CCIA database assembled by Value Health Sciences which compiles data based on the claims experience of several HMOs. It and similar compilations from other HMO databases have been used in several important studies of the efficacy of different treatments for common serious disorders. But these databases too have their problems. Samples are non-random, the information they contain is of uncertain reliability and, most importantly for present purposes, they do not capture much of the information that physicians use in deciding on treatments.

Hall *et al*, for example, artfully exploit data on 1220 patients treated for chronic heart failure by a medium-sized regional HMO and find that only 40% of the patients whose physicians are in the top 10% of physicians in their propensity to prescribe beta blockers had filled prescriptions for beta blockers. From these data it appears that even those physicians most prone to prescribe beta blockers often have reasons not to do so since less than half their patients appear to be taking them. Thus, the data suggest that physicians in the HMO community who fail to prescribe beta blockers are not guilty of improper medical management, even though best practice, based on randomized clinical trials, calls for their more widespread use. However, Hall *et al* recognize that not all patients prescribed beta blockers by their physicians will have filled their prescriptions or used HMO pharmacies if they did, so the 40% figure probably underestimates the proportion of cases where those physicians who are the most likely to believe beta blockers are the appropriate treatment for heart failure, prescribe them. More importantly, the HMO files contain no data identifying which heart patients were the best candidates for beta blocker use. The doctors in this sample may have prescribed beta blockers in virtually all of the cases in which they were likely to be helpful. A patient with symptoms that call

strongly for beta blocker use could justly complain if a jury received community standard data contaminated by many cases in which competent physicians would not have prescribed beta blockers. The standard the jury must apply is the standard that applies in cases like the plaintiff's. Problems are compounded when the physician community consists only of employees of a particular HMO, for their prescription behavior may reflect HMO cost containment pressures rather than their own judgments of the best feasible treatment. HMOs should not be able to shield their physicians from malpractice judgments by pressing for a lowest common denominator standard of practice. The community medical provider standard is justified in part by the assumption that it distills independent professional judgments of appropriate practice procedures. Within HMOs, independence is in some measure compromised.

These are not the only reasons to be wary of using pre-existing databases to derive community treatment standards for malpractice litigation. These data sets are used for important treatment efficacy and cost containment purposes. If those providing information to these databases knew their reports might figure in malpractice cases, the behavior of some respondents might be affected by this knowledge, perhaps diminishing the usefulness of their reports for their important primary purposes. In addition, standards of medical practice are constantly changing, perhaps more radically than ever with the advent of evidence-based medicine. Depending on the lag between alleged malpractice and trials, databases might provide a dated view of community practice standards. Beyond this, I am wary of creating a situation where the speed of adopting evidence-based procedures runs counter to professional interests in minimizing malpractice liability. As it stands now, malpractice litigation can be a force for positive medical change since experts often can be found who will testify that the community standard is to adopt best modern procedures. Whether or not this is true, the possibility that a jury will accept this standard is an incentive to keep up with changing best practices. Allowing proof of community standards by reference to practice procedures encapsulated in HMO files and other large databases will reduce and maybe even reverse this incentive.

A further set of quandaries arises from the relationship of patient characteristics to the treatments they receive. Hall *et al* tell us that in North Carolina radiation and chemotherapy following surgery for colon cancer are done far less frequently in patients 50 years of age and older than in

younger patients (18% vs. 47%). Does this mean that a 51 year old man with cancer whose disease reappeared after surgery would have no malpractice claim if follow-up chemotherapy more likely than not would have prevented the relapse, but a 49 year old man with the same history might successfully sue? If physicians treat diseased whites more aggressively than diseased blacks, will white patients be able to prevail on malpractice claims that blacks would lose? If one HMO is more tolerant of doctors who order expensive drugs and procedures than another, does a patient of the first HMO whose doctor chooses an ineffective, less expensive procedure have a malpractice claim that a patient of the second HMO would lack? These questions are rhetorical, I think. Surely we should not close our courts to socially disadvantaged groups because they are also disadvantaged in medical practice, nor do we want to reward those HMOs that offer their patients least by insulating them more from malpractice liability. It is hard to imagine a more perverse incentive system. Until malpractice law can come to grips with conundrums such as these, I would not encourage heavy reliance on HMO, Medicare and similar data sets as proof of community practice standards. The boundaries of both physician practice and patient communities are as yet too undefined.

Part III.

> *[T]here are three kinds of liars: the common liar, the d----d liar, and the scientific expert.*[20]

The case against using statistics to establish community practice standards would be stronger if it weren't for the alternative—expert witnesses. The problems of expert testimony are both notorious and real—so notorious that there is no need for me to rehearse them here.[21] A system which allows the parties to choose witnesses and to pay them well if they are willing to say what the party wants the jury

---

[20] Attributed to an anonymous lawyer in W. L. Foster, Expert Testimony: Prevalent Complaints and Proposed Remedies, 11 Harv. L. Rev. 169, 169 (1897).

[21] See S. Gross, *Expert Evidence*, 1991 Wis. L. Rev. 1113.

to hear is not well-designed to get at the truth.[22] The problem is compounded when experts give opinions rooted more in subjective experience than in well-established scientific theories or reliable empirical data, and when jurors' backgrounds do not help them identify the more accurate of two competing experts. The Meadow and Sunstein proposal and the proposals presented in this symposium are fueled not just by the perceived virtues of the empirical data the authors would have courts employ, but also, and perhaps primarily, by the perceived weaknesses of proving treatment standards through partisan expert testimony.

These proposals do not, however, eliminate expert witnesses, they simply replace or supplement one kind of expert witness with another. The various proposals for ascertaining community standards empirically are written as if good empirical research will reveal, even if imperfectly, the one *truth*. Yet the papers in this symposium suggest different ways of ascertaining community practice standards empirically, and very likely they will not yield the same truth.

In considering reform proposals, we should recognize that statistical evidence is socially constructed. Although the data an analyst reports may appear to be objective and clear in their implications, numbers can hide considerable disputable subjectivity. Research results are affected by the samples chosen, the variables considered, how these variables are operationally defined (including, in surveys, question wording and question order), and how analytic models are specified. One result of the malleability of statistical evidence is that when it is legally central, as it often is in antitrust, trademark infringement and sex discrimination litigation, it is common for fact-finders to hear competing statistical studies presented by competing experts. So long as the stakes are high enough, similar conflicts might be expected in malpractice litigation. Hence, if proposals like those made in this symposium are accepted, fact-finders often will still have to choose between the conflicting judgments of competing

[22] I do not necessarily mean to condemn the system from the point of view of truth-finding. Expert witnesses are a routine feature of criminal and civil litigation, and in most cases, it appears that judge and jury do tolerably well in sorting out the facts to get at the truth. It is not, however, clear whether this is because of how the system uses expert witnesses or despite it. No doubt, in some cases it is the one and in other cases the other. In still other cases, expert testimony contributes to unjust verdicts.

experts testifying to matters the jurors know little about. Litigation costs are, however, likely to increase, affecting settlement practices and perhaps increasing the dead weight losses of litigation.

The contrast between the working paper draft of Meadow and Sunstein's paper[23] and the published version[24] of this paper provides a simple yet revealing example of how the subjective judgments of statistical analysts can affect the results they present. In both papers, the average time to treatment in the Chicago emergency rooms they studied is 120 minutes, although the number appears to be a mean score in the draft version and a median in the published version. Yet the expert judgments appear less accurate in the published version (80 minutes for infectious disease [ID] specialists and 46 minutes for emergency room [ER] specialists) than in the draft version (87 minutes for ID specialists and 56 minutes for ER specialists). The reason for the discrepancy is that the authors report sample means in their draft version but medians in the published version.[25] There is nothing dishonest about either choice, but honest choices present different pictures and experts can be expected to make choices that present their side's case in the strongest possible light.

Also, the difference between expert judgments actual time to treatment in the two Chicago emergency rooms Meadow studied appear to be statistically insignificant in the draft version. (ID and ER expert judgments fall within the 95% confidence interval of actual times to treatment), but are reported as significant in the published version. The reason is that non-parametric statistical tests were used to analyze the data in the published version.[26] The authors appear to have a legitimate reason to prefer the non-parametric tests they used to a simple difference of means *t* test, but their choice was

---

[23] Meadow and Sustein, *supra* note 10.

[24] Meadow and Sustein, *supra* note 5.

[25] In Meadow's paper in this volume, medians are also reported, but they appear different from both prior reports. ID specialists report a median estimate of 1.4 hours, which equals 84 minutes, and ER experts report a median estimate of .9 hours which is 54 minutes. I have no idea of how the discrepancy arose, but if even medians cannot be reported consistently, it cautions us against taking the apparent objectivity of statistics at face value.

[26] Meadow and Sustein, *supra* note 5 at note 34.

just one of several they might have made to deal with non-normal distributions. Different choices might have yielded different results. Moreover, it is not clear that they would have found a statistical difference had they not chosen to compare expert judgments to data that combined results from the two ERs they studied. Although they offer a statistical justification for doing this (they cannot reject at the .05 level the hypothesis that times to treatment in the two ERs differed), the likely low power of their test may make its use disputable. Moreover, a case would arise in only one ER. Were an expert to testify to the normative time to treatment for doctors practicing in that ER, a good argument can be made that only data from that ER is relevant.

I don't mean by this discussion to criticize Meadow and Sunstein for any of the statistical choices they made on the road to publication, but their two papers illustrate how experts create statistics and the degree to which different legitimate expert choices can affect the picture that is finally presented. If this can happen when the same two people analyze the same data set with the same agenda using simple, standard statistics, imagine the variation that can arise and, indeed, the room for manipulation, when different experts, using more complex, less transparent methods, analyze possibly different data sets in order to prove opposing points.

Some might still prefer statistical experts to the experts who now testify, arguing that despite its flaws, statistical evidence is more *objective* than the statistically ungrounded opinions of hired experts. Given the hidden subjectivity in statistical analyses, I am unwilling to concede this point, yet even I share the intuition. In particular, when data have been collected and analyzed for another purpose, such as the data sets Hall *et al* describe, there may be little reason to suppose that errors in the data or the subjective choices that structured data collection are systematically biased to favor one side or the other in malpractice litigation. If it turns out that in case after case the data favor malpractice defendants, this is probably because actual practice standards are less protective of patients than experts testifying in good faith think. Also, statistical studies, whatever their flaws, must meet the test for relevance under FRE 401 and 402, and testimony based on them must meet the helpfulness test of FRE 702. Moreover, no one in this symposium is proposing that statistical evidence replace experience-based expert testimony; they are merely suggesting that it also be admissible. Indeed, it is hard to see how

statistical evidence can be excluded when expert testimony on practice standards is allowed. Surely some experts will supplement what they know from experience with what they learn from reported research, and if they do, they should be able to tell juries the bases of their opinions. It is also fair to cross-examine those who ignore published research and, if they are aware of it, to ask why it has not been influential. These are strong arguments for admitting the kinds of statistical evidence that participants in this symposium champion, but it is only the last I find convincing.

Although I will concede that statistical evidence collected for reasons other than litigation is likely to give a less biased portrait of practice standards than the opinions of competing experts, it may not provide a better view for reasons I outline in Part I above. Malpractice law's deference to community practice standards evolved in the context of how these standards could be proved. The standard did not contemplate proof so objective as to preclude juries from finding for plaintiffs when the truth was that local practice customs fell far short of widely known, locally feasible, best practices. Also, the unbiasedness of studies based on data collected for reasons other than litigation is easily exaggerated. While the data may not be biased by the prospect of litigation, analyses for trial might be. Moreover, if those furnishing data know that what they report may figure in litigation, their reports may be affected by this prospect, not only biasing the data for litigation uses, but also diminishing its value in its primary use.

I am also wary of the claim that statistical evidence will supplement but not supplant experience-based expert testimony. Surely the party favored by the statistical evidence will make a *Daubert* objection to the apparently more subjective opinion its opponent offers, and I expect that many courts will accept the invitation to exclude the latter. Expert opinion about community standards purports, after all, to predict what empirical investigation would find. If we have empirical data on community practice standards, how can conflicting opinion evidence help the jury? The problem is not that experience-based expertise is irrelevant, but rather lies in how *Daubert* and its progeny have modified FRE 402's command that all relevant evidence is admissible. Just as in many toxic tort cases, epidemiological evidence has come to occupy a position so privileged that it justifies excluding otherwise relevant in vitro studies and theory-based opinions, so I would expect statistical evidence

often to supplant rather than only supplement currently admissible opinion.[27] The result will most often be directed verdicts for defendants, because plaintiffs will lack admissible expert evidence to prove their claims about community practice standards.[28] Privileging statistical studies of community practice in this way would be a mistake, but it is easy to imagine it happening.

The strongest argument I see for admitting statistical evidence of community practices that honest experts will, and should be, influenced by extant statistical studies in forming their views about what is standard practice in particular medical communities. This does not mean their opinions need mirror the statistical results, for as experts they should be able to evaluate the strengths and weaknesses of relevant studies and to judge their implications for the case at hand. It would be artificial in the extreme to prevent these experts from referring to or being cross-examined on relevant statistical research. This can be done, however, without allowing the statistical research itself to be introduced into evidence. Although the distinction may appear bizarre to non-lawyers, the compromise is familiar when the law must deal with evidence that is, on the one hand, relevant and, on the other, likely to prejudice the jury, be overweighted by it, open up too many distracting issues, or take more time to present than it is worth. Thus, evidence of subsequent repairs, settlement offers, plea bargains and character may be introduced for some purposes but not others. Character evidence, when admissible, is often provable only by reputation evidence or opinion evidence rather than by more probative specific act evidence. Bad acts suggesting untruthfulness may be inquired into on cross-examination, but if a witness denies the act, evidence that it happened is not allowed.

---

[27] I am not saying that courts in toxic tort and product liability suits are wrong in privileging epidemiological evidence, but for reasons I give in Part II, I expect that statistical evidence of community standards will seldom have the methodological rigor or probative force of well-designed epidemiological studies. Moreover, I expect that statistically naive judges, especially judges sympathetic to defendants to begin with, will overweight the apparent objectivity of statistical evidence in deciding pretrial *Daubert* motions.

[28] This is not logically entailed, for the statistics might support the plaintiff rather than the defendant, but based on the papers presented in this symposium as well as likely settlement processes, I would expect that far more often than not, it will be defendants who benefit from statistical studies of community practice standards.

I suggest a similar compromise here. I would allow an expert witness on direct examination to refer to statistical research to support her opinion, but, as with reputation evidence of character, I would require the reference to be made briefly and in conclusory terms. For example, I would allow an expert witness to testify, "I believe that when a child with bacterial meningitis presents at an emergency room, the ordinary delay before the administration of antibiotics is about two hours, give or take thirty minutes, because that is what Meadow *et. al.* found when they studied two Chicago area emergency rooms." I would not allow the study itself to be introduced, nor would I allow the witness to describe it in detail as part of her direct examination. On cross-examination, opposing counsel would have the option of introducing the study into evidence or exploring its details with the witness in an attempt to discredit her opinion.

I would also allow statistical research to be used in cross-examining an expert whose opinion seems to conflict with the study's results, but as with the probing of bad acts to attack credibility, I would have the cross-examiner "concluded" by the witness's answers. This means the cross-examiner could not probe further if the expert did not know of the study, and the cross-examiner could not introduce the study or other evidence of it to refute a witness's explanation of why she wasn't influenced by it.

This compromise would limit the centrality of the statistical evidence at trial, preclude an additional, perhaps confusing, battle of statistical experts, substantially reduce incentives to conduct studies for the purpose of litigation and preclude courts from holding that statistical studies trump other expert opinion. At the same time, juries would be aware of relevant statistical research and could treat it as a factor when they weighed conflicting expert opinion. The limitations on the use of statistical evidence are not radical, but are like compromises evidence law often strikes between the full admissibility of evidence and its total exclusion. Indeed, my suggestions recall the Pre-Federal Rules treatment of most studies relied on by expert witnesses. The major difference lies in the degree to which my proposal limits the discussion of relevant empirical research on direct examination. An expert could describe a study's scope and report its bottom line, but not much more.

My problem with my proposal is that although it is easy to announce limits; it is much harder to

enforce them. Courts often use their discretion to admit relevant evidence, and even when they exceed what evidence law seemingly allows, appellate courts seldom reverse for such errors. Judicial frustration with conflicting expert testimony, the natural urge to explore in detail relevant evidence and the seeming objectivity of statistical studies are together a powerful force likely to erode the restraints I seek to impose on statistical evidence. My proposed policy would protect judges who appropriately limit statistical evidence from reversal, but I would not expect frequent appellate court reversals where limits were surpassed. Moreover, I would not be surprised if statistical evidence were unduly influential with some jurors, despite the truncated form I propose.

Even without these drawbacks, I regard my suggestion for dealing with statistical evidence of community practice as a second best solution. The best solution is to confront directly the issue that the symposium's advocates of statistical evidence bracket—the rule that physicians who perform like other physicians in their locality, region or nation are thereby shielded from malpractice liability, however short average behavior falls of best practice. I prefer a rule that holds physicians liable in malpractice whenever their behavior falls short of what might reasonably be expected of a competent physician in practice circumstances like the defendant's, and a patient is harmed because of it. If this were the standard and if it were not equated with what similarly situated physicians actually do, I would support the free admission of statistical evidence of community practice (subject to Federal Rules 401, 402, 403 and 702) as some, but not dispositive, evidence, of what a professionally competent physician in practice circumstances like the defendant's might do. Jurors could weigh empirical practice evidence along with other expert testimony, to decide what might reasonably be expected of a professionally competent physician. A jury might, for example, decide that a professionally competent emergency room will treat patients presenting with bacterial meningitis with the speed experts expect or as textbooks instruct rather than with the delay that empirical studies reveal. This might be because they perceive flaws in the studies or believe the defendant could reasonably be expected to perform at a higher level.

With the advent of evidence-based medicine, the promulgation of best practice standards and the accessibility of the information through the internet, I expect it will often be reasonable to hold

physicians to standards that the best among them meet, because all physicians have an obligation to comply with professional norms to "keep up." Lags in adaptation and skill development always exist, but tort law should provide incentives for shortening them rather than make empirical lags safe harbors for less than adequate care. Although there is seldom anything evil about physician mistakes, when patients are injured in treatment, there are always costs to be apportioned between physicians and health care institutions on the one hand and the patient on the other. A malpractice standard that apportions these costs to the health care sector when practice falls short of what a good physician (or health care delivery system) should do, seems reasonable to me, even if most physicians (or health care systems) are, at the time of the injury, not achieving this standard. My intuition is that the current standard with current modes of proof allows this nudge toward better practice.

In a world where we cannot prove precisely what most physicians (and health care systems) do, it appears that both experts and jurors think health care systems and professionals are performing better than they actually are. Empirical research, in the guise of better implementing current legal standards, promises to change what these standards, in practice, have been. Changing malpractice law to restore the living law's *status quo ante* is not a radical step, even if it appears to be a large one. Had the law actually followed the man on the Clapham Omnibus, it would have been shocked when he pulled out his cell phone and became so involved in an agitated conversion that he did not see a sidewalk barrier. All would have agreed that his behavior did not define reasonable care, although the empirical investigation into his behavior was impeccable. Similarly, impeccable empirical research into practice standards will not necessarily improve the current malpractice legal regime.