Chapter 26

LIABILITY FOR MEDICAL MALPRACTICE

PATRICIA M. DANZON
The Wharton School, University of Pennsylvania

Contents

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

Physicians are traditionally liable under a negligence rule of liability. Economic analysis of liability rules, including malpractice, assumes that the primary function of liability is injury prevention (deterrence). Compensation can be provided more efficiently through other forms of social or private insurance. In theory, a negligence rule creates incentives for efficient care, hence there should be no negligence, no claims and no demand for liability insurance. In practice, the incidence of negligent injury has been estimated at roughly one per hundred hospital admissions in the US and about one in seven physicians is sued per year.

These discrepancies between the theory and actual operation of the negligence system arise primarily because of imperfect information on the part of courts, doctors, patients, liability insurers and health insurers. Imperfect information and extensive health insurance lead to biased and uncertain legal standards. Uncertain legal standards create incentives for physicians to practice defensive medicine and incentives for plaintiffs and defendants to invest in litigation, leading to high overhead costs, such that compensation through the malpractice system carries a load of $1.50 per $1.00 of compensation.

Nevertheless, the extreme criticisms of the malpractice system are exaggerated. Malpractice premiums are less than 1 percent of total health care costs. There are no comprehensive estimates of defensive medicine costs; in any case such costs are likely to decline with the growth of managed care. Although claim disposition exhibits both Type 1 and Type 2 errors, negligent injuries are much more likely to lead to a claim being filed and payment to the plaintiff than non-negligent injuries, and awards are strongly related to loss incurred. The limited empirical evidence of provider response to liability and the deterrent effect of claims suggests – but cannot prove – that the net benefits of the malpractice system may plausibly be positive. Nevertheless, reforms designed to reduce inappropriate compensation and deter excessive litigation and defensive practice would make the system more cost-effective.

The empirical evidence, based primarily in the US, includes studies of malpractice injuries; physician response to liability; trends in claim frequency, severity (size), and claim disposition; and the malpractice insurance market. Analyses of actual and proposed reforms address tort reform, no fault, enterprise liability and optimal liability under managed care. More limited evidence is available on the negligence regimes in Canada and the UK, and the quasi no-fault regimes in Sweden and New Zealand.

Keywords

accidents, administrative costs, adverse events, awards, California malpractice study, caps on awards, Caveat Emptor, claim frequency, claim severity, collateral source offset, compensation, contingent fees, contractual approach, custom, defensive medicine, denial of coverage, deterrence, disposition of claims, economic theory of tort liability, efficient level of care, Employer Retirement Income Security Act (ERISA), English rule of fee shifting, enterprise liability, experience rating of liability insurance, fee for service, Harvard study, health care costs, imperfect information, injuries (iatrogenic), insurance, insurance crisis, learned hand standard, litigation costs, malpractice premiums, managed care, moral hazard, mutual, physician owned, negligence rule, New York malpractice study, New Zealand accident compensation, no fault, overhead costs, pain and suffering (non-monetary loss), periodic payment of awards, physician response to liability, policy issues, positive analysis, practice guidelines, prevention of injuries, regulation, scheduled awards, settlement out of court, sociolegal risk, standard of care, statute of limitations, strict liability, subrogation, Swedish patient compensation insurance, tort reform, uncertain standards, verdict, workers' compensation,

1. Introduction and overview

Although medical providers have been subject to liability for malpractice for centuries, malpractice only emerged as a major concern of physicians and of health policy over the last three decades. Since the late 1960s, malpractice claim frequency (claims per 100 physicians) and claim severity (average amount per paid claim) have risen dramatically but unevenly, leading to so-called "crises" in malpractice insurance markets in the mid 1970s and mid 1980s. Doctors in the highest risk specialties and regions, such as neurosurgeons or obstetricians in New York City, face annual premiums of $150,000–$200,000, although the national average is roughly $16,000 (weighted average over 20 specialties) [Bovbjerg (1995)]. Malpractice premiums account for roughly one percent of total health care spending, hence are not a significant contributor to the level or growth of health care costs. However, there remain no good empirical measures of the changes in medical care that are induced by liability – either cost-justified injury prevention that the system is intended to encourage or defensive practices that are not cost-justified and that are alleged to contribute significantly to total health care costs. In the 1970s and again in the 1980s, many states enacted changes in their laws governing legal liability and liability insurance. Federal legislation has been repeatedly debated but so far not enacted, reflecting widespread disagreement as to the nature of the problem, if any, and hence the desirability of change.

The traditional common law of medical malpractice holds health care providers liable for iatrogenic (medically-caused) injuries that are attributable to negligence. Under a negligence rule, the plaintiff must show that he or she suffered an injury; that the injury was caused by medical care; and that the provider's care deviated from due care. For medical liability, due care is defined as customary practice of practitioners in good standing or a significant minority of such professionals. Adverse outcomes that are within the normal risks of customary medical care are to be borne by the patient, possibly using first party insurance.

The theoretical analysis of medical malpractice liability is an extension of the economic analysis of tort liability more generally [Posner (1972), Brown (1973), Shavell (1980)].[1] In that analysis, the primary function of liability is to influence incentives for care and hence to reduce the rate of inappropriate accidents. Other potential functions of liability are recognized, in particular, providing compensation to injured patients, justice and retribution. However, compensation can be provided more efficiently through other forms of social or private insurance (see below), and economics has little to say about retribution. Economic analysis has therefore focused on the deterrence function of liability.

Liability rules are analyzed in terms of their ability to create incentives for first best efficient levels of care, that is, for investment in precautions to the point where the marginal cost is equal to the marginal benefit, in terms of reduction in expected injury

[1] Tort law refers to the set of legal rules and practices that govern wrongful injuries to person or property.

costs. Given certain assumptions, this analysis yields the important conclusions that, first, a negligence rule of liability can provide incentives for potential injurers to take first best efficient care, and second, a well-functioning negligence system should induce complete compliance, that is, there should be no negligence, no claims alleging negligence and no demand for liability insurance [Shavell (1982)]. A key necessary condition is that courts should define negligence as failure to take cost-justified precautions. In that case, it is always cheaper for a potential injurer to prevent an injury that would be deemed negligent than to insure against it.

The medical malpractice experience is seriously at odds with this prediction, of no negligent injuries and no claims. The incidence of negligent injury has been estimated at roughly one per hundred hospital admissions (Weiler, Hiatt et al. 1993). Whereas before the 1960s, only one in seven physicians had ever been sued, now about one in seven physicians is sued per year.

Economic theory suggests that the discrepancies between the negligence system in theory and its operation in practice arise primarily because the decision-makers – courts, doctors, patients, liability insurers and health insurers – lack the information that is assumed by the models (Danzon 1991).

Imperfect and asymmetric information can lead to legal standards of care that are systematically biased and have high variance. Legal standards that are unpredictable and open to influence can create incentives for physicians to practice defensive medicine and incentives for plaintiffs and defendants to invest in litigation to influence the outcome, leading to high overhead costs. Roughly forty cents of the malpractice insurance premium dollar is spent on litigation, 20 cents is insurance overhead and only forty cents reaches the plaintiff as compensation. Whether the deterrence benefits of the malpractice system are sufficient to outweigh its high administrative costs plus any net costs of defensive medicine remains a critical but unresolved question.

The economic literature on medical malpractice falls into two broad categories. The largest category is primarily empirical and positive in focus, aimed at providing evidence on the actual operation of the malpractice system. It includes studies that attempt to measure the number of malpractice injuries and claims; physician and patient response to liability; determinants of the number and size of claims, claim disposition and other issues related to the legal system; and the supply and pricing of malpractice insurance. A second, smaller literature is directly concerned with the policy issues of malpractice reform, including evaluation of actual and proposed changes. Most of the literature draws on the US experience, which has provoked the fiercest policy debate and generated the most reliable data and analysis. However Canada and the UK have also experienced rising claims and premium costs in the 1980s, giving rise to proposals for policy changes [Dewees, Coyte et al. (1989), Fenn (1993), Towse and Danzon (1998)]. In the early 1970s, Sweden and New Zealand replaced their traditional negligence systems with quasi no-fault systems of liability, and this approach has been considered as a possible model in the US, the UK and Canada.

In this Chapter, Section 2 reviews the standard economic analysis of tort liability and identifies the problems in applying it to medical care. Section 3 reviews empirical evi-

dence on the incidence of medical injuries, trends in malpractice claims and claim disposition. Section 4 discusses malpractice insurance contracts and market performance. Section 5 reviews the evidence on the effects of liability on medical practice, including defensive medicine. Section 6 summarizes the evidence on the overall costs and benefits of the current system, and the case for reform. Sections 7 and 8, respectively, evaluate the theory and evidence on traditional tort "reforms" (such as caps on awards) and more radical alternatives such as no-fault plans and contracting out of liability. Section 9 reviews the emerging issues and challenges posed by managed care. Section 10 reviews the experience in the UK, Sweden and New Zealand; Section 11 concludes.

2. The theory of tort liability

2.1. Tort liability with perfect information

Accidents are a costly by-product of other beneficial activities, for example, medical care or driving a car. The probability and size of losses can be reduced at a cost, either by reducing the rate of risky activities (number of surgeries or miles driven per year) or by taking costly precautions per unit of activity (care per surgery or per mile driven). Brown (1973) showed that safety can be viewed as a good like any other. Efficient investment in safety requires producing both the efficient level of safety and using the lowest cost mix of inputs. The problem can be viewed most narrowly as minimizing the total social cost of accidents, defined to include the cost of precautions and the cost of injuries. If insurance is available and claims disposition is costly, the maximand is expanded to reflect the utility cost of injuries, net of optimal compensation and transactions cost. An even broader formulation maximizes a social welfare function that reflects the utility value of the beneficial activities of which accidents are a by-product, in addition to the costs of injuries, prevention and overhead (Shavell 1980).

In these economic models, alternative liability rules allocate the costs of accidents in ways that affect private incentives for investment in safety – the "deterrence" function of liability. The main rules under consideration are no liability (caveat emptor), negligence and strict liability. Efficient rules are those that create incentives for socially optimal care by all participants. Tort liability also provides a source of insurance and compensation to injury victims. Although the compensation function of liability features prominently in the policy debate, economics has little to say about the equity of different compensation rules. Economic analysis has been used to evaluate the efficiency of tort compensation structure and amounts (Section 7) and the efficiency of providing compensation through tort liability vs. either private or public first party insurance. On the latter point, the conclusion is clear: liability is a much more costly insurance mechanism than first party insurance (Section 6). The implication is that if liability is to be justified as an efficient institution for dealing with risk, its deterrence benefits must outweigh its added costs.

In analyzing the deterrent effects of different legal rules, it is useful to categorize accidents as unilateral (optimal care is positive for just one party) or bilateral (optimal

care is positive for both the injurer(s) and the victim). Medical injuries are generally assumed to be unilateral, due to asymmetric information between providers and patients. A second useful distinction is whether accidents occur between strangers (for example, most automobile accidents) or between parties who are in a market relationship, such as the patient-physician relationship for medical injuries, the employer–employee relationship for workplace injuries or the consumer–producer relationship for product-related injuries. Shavell (1980) provides a systematic analysis of the incentives for care in these different contexts – unilateral or bilateral accidents, between strangers or market participants – under alternative liability rules (including no liability, a negligence rule and strict liability) and under different assumptions about information asymmetry. The Appendix applies a similar model to medical malpractice.

For accidents between strangers in a non-market context, an accident is like any externality: the injurer has no incentive to invest in any care in the absence of liability or regulation. Modeling accidents between strangers as a non-cooperative game, Brown (1973) shows that a negligence liability rule creates incentives for efficient care, provided that due care is defined by the "incremental Learned Hand standard". According to this standard, defined by Judge Learned Hand, negligence occurs if "the loss caused by the accident, multiplied by the probability of the accident's occurring exceeds the burden of the precautions the defendant might have taken to avert it".[2] In other words, negligence consists of failure to take precautions if their cost is less than the expected damages averted. If defined in terms of marginal cost and benefit, this defines negligence as failure to take efficient precautions.

However, if the potential injurer and victim are in a contractual relationship, as in the physician-patent relationship in medical care, the value of liability depends on the extent of information [Spence (1977), Shavell (1980)]. If customers are fully informed, their valuation of safety is appropriately internalized to producers through the prices that they are willing to pay for safer products. But if customers misperceive risks, a producer who invests to make a product safer cannot recoup the investment through a higher price. In the medical context, if the liability rule is *caveat emptor* and if patients underestimate surgical risks in general and cannot monitor the individual surgeon's care, there would be too many surgeries and too little care per surgery. Conversely, there would be too little risk-taking if consumers overestimate risks. Since many medical services are infrequently purchased, it is difficult to become an informed shopper until it is too late. The existence of a core of informed shoppers will not necessarily assure appropriate quality for everyone, because the product is patient-specific, not a mass-produced, homogeneous good.

In theory, a well-functioning negligence rule could, given certain assumptions, create incentives for optimal care per procedure. To achieve the optimal rate of risky procedures, the definition of negligence should include liability for performing "unnecessary" procedures, if patients misperceive average risk. But the simple model of negligence liability as an efficient system of deterrence assumes perfect information on the part of

[2] US vs. Carroll Towing 159 Federal Reporter 2d 169 (1947).

courts, potential injurers and, ex post, of victims. The model assumes that courts costlessly enforce the efficient standard of care, providers know the due care standard, and that patients file a claim if and only if they are injured due to negligence. Under a perfectly functioning negligence rule there would be no negligence and no claims, since by definition it is cheaper to prevent injuries that would be deemed negligent than to pay for the resulting damages [Shavell (1982)]. There would also be no defensive medicine. But with perfect information, there would be no need for negligence liability. More realistically, consumer information in medical markets is imperfect, which creates a prima facie case for provider liability. Whether in practice exposing medical providers to tort liability improves efficiency depends on the information available to decision-makers under such a rule. If doctors, patients, courts and liability insurers lack good information about appropriate medical care and legal rules, then the operation of the negligence system in practice can diverge significantly from the theoretical ideal.

2.2. Tort liability with imperfect information

2.2.1. Bias in the custom-based standard of care

For professional liability of physicians, attorneys, architects and other professionals, courts generally define due care as customary practice of practitioners in good standing, acknowledging their own imperfect information. The explicit and independent use of a Learned Hand cost-benefit calculus to define negligence is extremely rare.[3] By deferring to custom, courts forego any attempt to correct any systematic bias in customary care that results from consumer misperceptions in medical markets. Yet the correction of such distortions is the main justification for liability assumed in the economic literature. A custom-based standard is likely to be systematically biased, relative to first best efficient care, in precisely those circumstances where market failure is significant. At best, a custom-based negligence rule could prevent significant deviations from the (non-optimal) norms that consumers have come to expect.

The prevalence of comprehensive private or public insurance for medical care creates further systematic bias in customary care. In the US, over 80 percent of the population has some form of health insurance, typically with modest co-payment and premiums unrelated to own use (experience rating). Most other countries have virtually universal systems of public or quasi-public insurance, with even lower co-payments. Insurance that reduces the point-of-purchase price of medical care to patients creates moral hazard [Pauly (1968)], in the absence of other provider incentives or controls by insurers. Norms of quantity and multiple dimensions of quality are likely to be distorted, relative to the first best optimum.

The nature of the biases in customary care depends on the structure of insurance coverage, in particular, the financial incentives created by provider reimbursement and

[3] Helling vs. Carey 83 Wash. 2d 514, 519 P2d 981, 983 (1974).

patient co-payments. Under traditional fee-for-service insurance with low co-payments for patients, moral hazard leads patients to prefer overuse of costly procedures, relative to a first best optimum. Providers acting as good agents for patients have every incentive to comply. On the other hand, strict fee regulation, as tends to occur in public systems, creates incentives for physicians to reduce real resource intensity per encounter, leading to a high frequency of very short visits – as, for example, in Germany, Japan and traditional US Medicaid. Thus fee-for-service reimbursement leads to a high volume of reimbursable encounters and procedures, increasing the likelihood of errors of commission. Errors of omission are also possible, however, particularly if fees per encounter/procedure are regulated at low levels, since this creates incentives for physicians to reduce resource input per encounter. Skimping on time and cognitive effort can lead to errors in diagnosis and treatment.

Capitation forms of payment are increasingly being adopted under managed care in the US and in some other countries, particularly for primary care physicians, to correct the incentives for overuse created by fee-for-service reimbursement. Since the capitated physician faces a positive marginal cost but receives zero marginal revenue per unit of additional service or effort, capitation may create incentives for suboptimal quantity and quality of care if patients have imperfect information about quality or face costs of switching physicians. The deterrence value of liability may therefore in theory be greater under capitation than fee-for-service [Danzon (1994b)], particularly for deterring errors of omission. Note that defining an appropriate due care standard may be particularly difficult for courts at a time of transition from fee-for-service to capitation, as is occurring in the US. If the courts continue to define due care in terms of (presumably excessive) customary practice that developed under fee-for-service insurance, liability will undermine the potential efficiency gains from moving to managed care. Drawing the line between appropriate reductions in excess, which is the promise of managed care, and inappropriate underprovision, which is its potential downside, may require resort to an external standard other than customary care [Danzon (1997) and Section 9 below].

Thus when both patients and insurers lack the information necessary to determine appropriate care and to monitor the care actually delivered, the safe conclusion is that customary care is likely to deviate systematically from optimal care, both in quantity and quality. However, the direction and extent of bias depends on the level and form of provider reimbursement, professional norms, the effectiveness of surrogate monitoring of quality and outcomes, and other factors.

An alternative benchmark would recognize that the full-information, first best optimum is not a relevant standard, given the costs of obtaining information and controlling moral hazard in health insurance markets. The issue is then whether a liability standard defined in terms of customary care might be second best optimal, given the information available in insurance markets. However, US markets for health insurance and hence medical care are distorted by the tax subsidy to employer contributions to private insurance and direct subsidies to public insurance Medicare and Medicaid. Subsidies to health insurance drive a wedge between private and social costs. Medical norms will

tend to reflect the distorted private costs, assuming that physicians act as reasonably good agents due to professional norms and competition in medical markets. Managed care is no antidote to subsidy-induced distortions since it is equally subsidized.

Thus in the case of medical care, the standard model of potential efficiency gains from tort liability is at best an approximation. Liability may help enforce that actual care conforms to customary norms which themselves are non-optimal due to imperfect information and subsidies to insurance. If courts were to attempt to weigh social costs and benefits, this would create severe conflict between legal standards and competitive pressures in medical care markets that are driven by private costs and benefits, at least in the US. This tension may be less in a system with public provision such as the UK NHS, if doctors adopted treatment norms based on social rather than private costs and benefits.

2.2.2. Uncertain medical and legal standards

The literature on variation in medical practice norms within countries [Phelps (2000)] concludes that much of this variation reflects differences in doctors' beliefs about appropriate care. Courts accommodate some difference in medical opinion (and possibly in patient preferences) by recognizing "customary practice of professionals in good standing or a significant minority of such professionals". A legal standard that recognizes a range of acceptable medical practice is potentially optimal, given uncertainty of medical opinion and differences in consumers' preferences and willingness to pay for care. However, this necessarily contributes to variance in judicial decisions. A further implication of a custom-based standard is that appropriate care in each case must be defined by medical experts. Although such experts owe a duty to the court, they are paid by the litigants and are presumably selected to try to influence the outcome, contributing further to uncertainty.

Craswell and Calfee (1986) show that a negligence rule with an uncertain standard is likely to create nonoptimal deterrence incentives. The precise effect depends on the bias and the variance of legal standards relative to efficient care. In this context, there is some presumption for excessive care. With uncertainty, the physician cannot be sure of avoiding liability simply by taking the required level of care. Because liability is all-or-nothing, by incurring a small additional cost his probability of a large penalty may be significantly reduced. Thus uncertain legal rules may be a significant factor contributing to incentives for defensive medicine.

Imperfect information about medical and legal standards contributes to errors by patients and their attorneys in filing claims. The evidence from several studies in the US indicates that many valid claims are not filed and many invalid claims are filed (Section 3). Only 43 percent of claims filed with insurance companies receive any payment [Bovbjerg (1995) and sources cited therein]; the remainder are either dismissed or dropped. Although negligent injuries are more likely than nonnegligent injuries to lead to claims and valid claims are more likely than invalid claims to receive payment [White (1994)], the system is far from perfect. The more variable are legal rulings, the

greater the incentive for plaintiff and defense to invest in legal expense to influence the outcome.

From the standpoint of physicians, invalid claims and uncertain legal standards generate a demand for comprehensive liability insurance, including legal defense insurance. This contradicts the theoretical prediction (which assumes perfect information and efficient standards), that there should be no demand for liability insurance because it is always cheaper to be non-negligent. Liability insurance would not interfere with deterrence if it were perfectly experience-rated, with premiums adjusted to reflect the actuarial risk implied by the physician's actual level of care. In practice, however, malpractice insurance premiums are based primarily on location, medical specialty and limits of coverage; individual experience-rated adjustments are infrequent and rarely based simply on the number of claims filed or paid. Ellis, Gallup and McGuire (1990) have shown that experience-rating based on a simple count of claims filed or paid would expose providers to significant risk of inappropriate surcharges, because of the large number of false positive claims [see also Sloan, Mergenhagen et al. (1989)].

Nevertheless, the widespread purchase of liability insurance with minimal experience-rating does not necessarily imply that deterrence incentives are nonexistent or even suboptimal, as some have argued. Physicians with consistently bad claims experience, adjusted for specialty, may face a surcharge on their premium, restrictions on their scope of covered practice (for example, no surgery) or be denied coverage by low cost insurers. However, an important implication of pervasive liability insurance with no formal coinsurance and minimal experience-rating, is that deterrence probably derives more from the uninsured time, embarrassment and reputation costs faced by physicians, which plausibly depend more on claim frequency of claims than on size of awards. This has implications in evaluating reform proposals, as discussed below.

2.3. Other quality control mechanisms

Liability is only one of several mechanisms that may correct the distortions that result from asymmetric information in medical markets. Altruism, professional or ethical concerns may motivate physicians to act as better agents for patients than would be predicted by models that assume purely self-interested income maximization (Danzon 1994b). In most countries, other regulatory mechanisms such as state licensure and boards of medical quality assurance provide coarse screens to eliminate persistent incompetence or misconduct by medical providers, although in practice implementation may be limited [Sloan, Mergenhagen et al. (1989)]. Indirect market mechanisms such as hospital credentialing committees, reputation and referral networks may provide some substitute for informed consumers. Performance measurement, to permit monitoring of providers by patients, employers and other third party payers through outcomes reporting and other quality indicators, is a rapidly growing field in the US, the UK and other countries [for example, Eddy (1998)]. Since professional ethics, regulation and indirect market forces all provide some correction for the market failure that results from imperfectly informed consumers, liability should only be used if it is cost-justified (benefits

exceed costs) and cost-effective relative to these alternative corrective mechanisms [see Spence (1977) and Appendix].

3. Empirical evidence on injuries and claims

3.1. Adverse events and negligent injuries

The theoretical assumption, that an accident is a well-defined event, is reasonable for automobile, product and acute workplace injuries that can be causally related to specific activities. However, a patient who seeks medical care is already either ill or injured and medical care cannot guarantee perfect health and immortality. Thus defining a medical "injury" and a causal relation between an imperfect medical outcome and the medical treatment is problematic. The amount and quality of medical services shifts the probability distribution of outcomes, hopefully with more weight for good outcomes but possibly also increased risk of some bad outcomes. For example, surgery reduces the risk of death from appendicitis but entails risk of other infections or complications, even if competently performed. Identifying a medical injury therefore requires an assumption about the expected distribution of outcomes, which presupposes an assumption, implicit or explicit, about the expected or appropriate level of medical care. Defining a negligent injury requires specifying a particular threshold within the range of expected care. Given the complexity of defining and recognizing medical injuries and negligent injuries, it is not surprising that comprehensive data on the number of iatrogenic injuries, with or without negligence, are not routinely collected.

The best available information is from two surveys of medical records of hospitalized patients, the first in California in 1974 [Mills (1977), hereafter California study] and the second in New York in 1984 [Weiler et al. (1993), hereafter Harvard Study]. In both studies, a large sample of hospital records was reviewed by experts in law and medicine – 20,864 patients in 23 hospitals in California and 31,429 patients in 51 randomly selected hospitals in New York – to identify the incidence of injury due to medical care and the subset of these injuries caused by negligence on the part of a health care provider. The California study concluded that 4.65 percent of hospitalized patients suffered an injury due to medical care; of these, 17 percent (or one in 126 patients) involved negligent injury. In the New York study, 3.7 percent of patients suffered an injury due to medical care and of these, 28 percent were attributed to negligence, that is, roughly 1 percent of all hospitalized patients suffered a negligent injury. If extrapolated to all New York hospital discharges in 1984, this would imply that 98,610 hospitalized patients per year suffered a medical injury. Of these, 57 percent were minimal with recovery in one month, and an additional 14 percent lasted less than 6 months. However, over 6,000 (7%) patients suffered permanent disabilities and another 13,400 (14%) died. If extrapolated to the US population as a whole, this would imply 150,000 iatrogenic fatalities annually, more than half of which may be attributed to negligence. This far exceeds the rate of motor vehicle fatalities (50,000 deaths per year) and occupation related fatalities

(6,000 per year). These studies and the resulting estimates omit injuries that occur in ambulatory settings, unless they resulted in hospitalization.

However, these startling figures on iatrogenic injury rates should be viewed with caution. As the authors note, a substantial (but undetermined) proportion of the patients whose deaths were attributed to medical management were seriously ill and many would have died from their underlying illness in months, days or even hours, whereas most victims of automobile or workplace injuries are healthy. Second, this apparently high rate of iatrogenic injury in part reflects the broad definition of injury used in both studies. The Harvard study defined an iatrogenic injury as "any disability caused by medical management that prolonged the hospital stay by at least one day or persisted beyond the patient's release from hospital".[4] Thus the count of injuries implicitly reflects the standard of care considered appropriate by the analysts, including their expectation of an appropriate length of stay and reasonable medical outcome. Any (negative) deviation from the expected outcome is considered an injury.

A negligent adverse event was defined as the consequence of treatment that failed to meet the standard of the average medical practitioner. The Harvard study used a modified locality standard "similar to that employed by peer review organizations ... we did not require that a reviewer imagine that all the resources of, say, a tertiary care teaching hospital were available at small community hospitals" (p. 35). Similarly, the California study defined negligence in terms of the standards likely to be applied by a jury. Since neither study attempted to define negligence by weighing marginal costs and benefits of additional precautions, the resulting count of "negligent injuries" does not necessarily correspond to the number of economically inappropriate injuries, for which costs of precautions are less than expected damage costs.

As in the California study, the Harvard study found that the percentage of the injuries attributed to negligence was higher for serious injuries than for minor injuries – for example, 51 percent of deaths were attributed to negligence, compared to 23 percent of impairments lasting less than 6 months. While it is possible that injuries caused by negligence have systematically worse outcomes than injuries not involving negligence, this finding could also be influenced by the definitions. Specifically, the definition of negligence, which invokes legal standards, may implicitly set a higher threshold of injury severity than that required for an adverse outcome to count as a nonnegligent injury.

These US-based studies of iatrogenic injury have been used to estimate injury rates for other countries [for example, Smith (1990)]. Such extrapolations are problematic because of cross-country differences in standards of care, thresholds of legal negligence and – presumably – in rates of iatrogenic and negligent injury. The direction of bias is

[4] Weiler et al. (1993, p. 35) refer to "those adverse events that were the unintended or unexpected harmful consequences of medical intervention, and that prolonged the hospitalization beyond the time required by the underlying illness and/or caused disability at the time of hospital discharge or death". This would exclude adverse but inevitable consequences of appropriate treatment, such as amputation of a limb to treat bone cancer.

uncertain a priori. The higher claim frequency and premium cost in the US should create stronger deterrence incentives for US doctors and hospitals. Financial and agency incentives may also differ. More fundamentally, any attempt to compare rates of adverse events and negligent injuries must specify the standard of care implicit in defining injuries and negligence, and whether any observed differences in estimated injury rates in fact reflect differences in care patterns that are appropriate, given differences in costs, preferences and other factors, or whether they reflect true differences in rates of inappropriate injury, conditional on different standards.

3.1.1. Causes of medical injuries

The causes of malpractice injuries are not well understood but both patient and provider characteristics play a role. In the Harvard study, age over 45 and Medicaid payer status were associated with significantly higher rates of adverse events; controlling for these and other factors, race and gender were not significant.[5] Age over 65 was also associated with significantly higher percent of adverse events due to negligence. Weiler et al. comment that "although their higher injury rates can be attributed in part to the frail physical state of older patients, the higher negligent injury rates may also be the result of the quality of care the elderly receive from health care providers". It may also reflect the fact that the definitions of adverse events and negligence are not age-specific. By contrast, economic definitions of optimal injury rates would take into account the age/condition-specific distributions of expected outcomes and costs of preventing adverse outcomes. If the costs of preventing mishaps increase with age and severity and/or the expected benefits to precautions decline, then optimal injury rates would increase with age and illness severity.

For purposes of prevention, an important issue is whether negligent injuries are caused largely by occasional inadvertent lapses of many, normally competent providers or by a minority of incompetent, physicians and low quality hospitals. (Of course this assumes, following standard legal practice, that all providers should be held to a common standard of care, whereas if costs and prices vary, standards of care would optimally differ.) Studies consistently show that both factors play a role. For the hospitals in the Harvard study, the percent of adverse events due to negligence averaged 25 percent, but ranged across hospitals from 1 percent to 60 percent. Controlling for patient age and DRG category (a proxy for illness severity), the factors associated with relatively high rates of adverse events were university teaching status, urban location, and small or medium size (relative to large). Factors associated with a high proportion of injuries due to negligence were high proportion of minority discharges (positive), university teaching and proprietary status (negative). Resources, case mix, location and related factors thus appear to affect outcomes.

[5] Low income was significantly negatively related to adverse event rates; however this is presumably positively correlated with Medicaid payer status which is positively related to adverse event rates [Weiler et al. (1993, p. 48)].

Similarly, analysis of claims experience of groups of physicians indicate that, after controlling for medical specialty, the distribution of claim frequency is more concentrated than would be expected based on chance alone [Rolph (1981), Nye and Hofflander (1987), Sloan, Mergenhagen et al. (1989), Ellis, Gallup et al. (1990)], as discussed further in Section 4.1.

In the Harvard study, diagnostic errors accounted for only 8 percent of adverse events, but 75 percent of these were attributed to negligence. Drug-related events accounted for 19 percent of all adverse events, but these were less frequently attributed to negligence than most other categories. Note that despite the Harvard study's initial sample of over 30,000 records, there are only 1,278 adverse events and 306 negligent adverse events. This small sample limits the scope for detailed analysis of causation. Identifying the medical interventions that are most at risk of resulting in negligent injury would require, in addition, data on the underlying frequency of treatments, which was not part of the study.

3.2. Malpractice claims vs. negligent injuries

The simple theoretical models of tort liability assume that a claim is filed if and only if a negligent injury occurs. In reality, several studies suggest that both false negatives (failure to file valid claims) and false positives (claims filed without negligent injury) occur frequently in medical care.

The California study did not directly compare claims to injuries. However, a comparison of the number of malpractice claims filed in California relative to the estimated number of negligent injuries concludes that at most one in ten of negligently injured patients filed a claim, assuming no false positive claims. Only 40 percent of these claimants received compensation through the tort system (Danzon 1985a). The Harvard study directly compared filed claims to identified injuries. The total number of malpractice claims was only about 15 percent of the number of negligent injuries.[6] However, only 2 percent of patients identified as having sustained an injury due to negligence filed a claim. This suggests a large number of both false positive and negative claims, at least relative to the definition of negligence and the data available to this study. The ratio of claims to negligent injuries was much higher for serious injuries: roughly one claim was filed for every three such injuries and one in six was paid [Weiler, Hiatt et al. (1993)]. The California data also indicated a lower ratio of claims to injuries for minor compared to major injuries, and lower for persons over 65. These findings are consistent with a simple economic model of the decision to file a claim, with fixed costs of filing and a payoff that depends on economic loss, hence increases with injury severity and decreases with patient age.

[6] The definition of a claim includes claims filed by patients with insurance companies, even if no legal suit was filed. Multiple claims per incident are aggregated. In the Harvard study, there were roughly 1.5 provider claims per patient incident.

3.3. Trends in malpractice claims

Several studies have examined the determinants of trends in claims over time and persistent differences across states and countries. Although malpractice liability has existed for centuries, such actions were rare until the late 1960s. In the US from the late 1960s to the mid-1980s malpractice claim frequency (number of claims per 100 physicians) increased at more than 10 percent a year. Claim severity (the average payment per claim paid) increased at roughly twice the rate of general inflation [Danzon (1986)], with some evidence of disproportionate growth for the highest stakes cases [Shanley and Peterson (1987)]. Claim frequency reached a peak of about 17 claims per 100 physicians in 1986, and then stabilized around 15 claims per 100 physicians per year.[7] Failure to anticipate the surge in claims precipitated "crises" in liability insurance markets in the mid-1970s and mid-1980s.

A simple economic model views the decision to file a claim as an investment with an uncertain payoff that depends on the nature of the injury, the legal rules defining negligence and compensable damages, and the costs of filing [Danzon (1984a)]. This model implies a simple econometric model in which the frequency of claims per capita, at the state level, depends on the frequency and characteristics of medical treatment; the legal rules that affect probability of winning and expected award, and the costs of legal input.

The empirical evidence confirms that the growth in claim frequency in the 1960s and early 1970s was significantly related to the increase in surgical interventions which increased the number of adverse outcomes that could be attributed to medical care. At the same time, pro-plaintiff shifts in legal doctrine increased the expected payoff to filing a claim, by increasing the grounds or reducing the cost of a showing of negligence. For example, the abolition of the locality rule substituted a statewide or national standard for a local standard of due care, which plausibly increased the number of injuries that would be deemed negligent. Perhaps more important, the move to a national standard meant that out-of-state experts could testify as to the standard of care, which allegedly broke down the "conspiracy of silence" that prevailed when local experts were required to testify. The abolition of charitable and government immunity exposed voluntary and government hospitals to suit. The doctrine of *respondeat superior* extended the liability of hospitals for actions of their employees and staff physicians; and requirement to obtain the patient's informed consent to treatment was increasingly defined in terms of the information that a reasonable patient would want, rather than what was customary for physicians to provide. But these factors had run their course by the mid-1970s and cannot explain claim growth in the 1980s.

In response to the malpractice insurance crisis of the mid-1970s and again in the 1980s, many states enacted one or more tort reforms.[8] For example, 18 states adopted

[7] These figures from Bovbjerg (1995) are based on the experience of The St. Paul Fire and Marine Insurance Company, which is the largest writer of malpractice insurance, operating in 42–43 states. More recent data are not maintained on the same basis. The experience of other insurers may differ.

[8] Tort law and insurance regulation are traditionally areas of state jurisdiction.

caps on awards or collateral source offset before 1985 and 23 states adopted these reforms between 1985 and 1990 [Kessler and McClellan (1996)].[9] Estimating the effects of these changes is problematic because each state's legislation is unique to a degree. Moreover, reforms were often enacted in response to "crisis" conditions, which raises the possibility that estimates of effects of reforms may be biased due to endogeneity and other, unobserved state characteristics. Nevertheless, the findings from three empirical studies that use data from 1975–1984 are reasonably consistent [Danzon (1984a, 1986), Zuckerman, Bovbjerg et al. (1990)]. Caps and collateral source offset reduced claim severity by 19–39 percent and 11–50 percent, respectively, relative to what it would have been in the absence of the reforms. Shorter statutes of limitations reduced claim frequency – one year off the statute of limitations for adults is estimated to reduce claim frequency by 8 percent. Collateral source offset has also reduced claim frequency (by 14 percent) presumably because of the feedback effect from lower expected awards to reduced incentive to file claims. Other reforms do not seem to have had significant effects.

Although medical and legal factors account for some of the trends and interstate differences in claim frequency and severity, much remains unexplained. The similarity of claims filed per negligent injury in the California and Harvard studies suggests that claim growth in the intervening years cannot simply be attributed to the "catch-up" filing of claims for a larger percentage of negligent injuries. Growth in the number of lawyers per capita is not a statistically significant explanatory factor, after controlling for other attributes of lawyer-dense areas. Danzon (1984a) found that urban areas tend to have much higher claim frequency and severity, but that this urban phenomenon could not be explained by specific observable characteristics of urban areas such as income, unemployment, welfare recipiency or population turnover rates (intended as a proxy for the "physician–patient relationship"). The growth in malpractice litigation coincided with abnormal growth in other areas of tort law, notably product liability, suggesting a role of common legal or social factors. However, the growth in product claims can plausibly be attributed to pro-plaintiff changes in product liability rules [see, for example, Henderson (1988)], whereas no comparable legal changes occurred for medical cases. It is possible that numerous subtle changes in case law and rules of evidence made it easier for a plaintiff to establish a cause of action and get to a jury, which effectively reduces the expected cost and increases incentives to file marginal claims. Such hypotheses are plausible but very hard to test.

An interesting but unanswered question is whether the increase in claim frequency is largely a response to higher expected awards, leaving the increase in awards as the main factor to be explained. Evidence from workers' compensation [for example, Butler (1983)] shows that claim rates respond positively to benefit levels. A similar supply

[9] Under traditional damages rules, the tort award is not reduced by the amount of compensation available to the plaintiff from other (collateral) sources. Collateral offset laws provide for the tort award to be reduced for some forms of private and public insurance.

response is plausible for medical malpractice, although much harder to measure because malpractice payment levels are not statutorily determined; rather the average observed award is endogenous, reflecting the mix of cases actually filed. In a simple model of the settlement process, the settlement amount depends on the probability of plaintiff verdict (p) which depends on the degree of negligence N; the size of verdict (V) if successful which depends on damages (D) and rules of compensation; the litigation costs of going to verdict for defense and plaintiff (C_d, C_p); and a bargaining parameter (g):

$$S = p(N)V(D) + (C_d - C_p)^g. \tag{3.1}$$

Thus the observed mean payment S_x reflects not only the legal rules of damages $V(D)$ but also the distribution of damages and plaintiffs' probability of winning, litigation costs, all of which are endogenous, depending on the distribution of cases filed.

The rate of growth of malpractice claim frequency and severity was as high in Canada and the United Kingdom in the 1980s as in the United States [Dewees, Coyte et al. (1989), Danzon (1990a, 1990b)]. But in 1987 physicians in the United States were still at least 5 times more likely to be sued than physicians in Canada or the UK. The mean payment was somewhat higher in the US than in Canada. However, the similarity of mean payments does not permit inferences about cross-country differences in net compensation for the same injury for several reasons. First, US awards generally include past and future medical expenses, whereas tort awards in the UK and Canada are net of medical costs that are borne by public health care systems. Second, in the US, the reported award is gross of the attorney's contingent fee (typically one third), whereas the UK and most Canadian provinces apply the English rule, that the loser pays costs. Third, the mean observed payment depends on the actual mix of cases filed and the incentives to settle out of court, all of which may differ across countries. In particular, to the extent that the higher frequency of claims in the US reflects disproportionately more minor injuries or cases of dubious merit (low pV) which settle for a relatively low payment, the mean observed payment in the US provides a downward biased estimate for the expected compensation for the case mix in Canada or the UK, where the absence of contingent fees is likely to deter more cases with low damages or low probability of winning. It is interesting that the growth in malpractice claim costs has occurred in Canada and the UK despite the fact that these countries have limits on awards for pain and suffering, have lower medical costs and lower rates of growth of medical costs, and do not permit contingent fees for plaintiff attorneys – all factors commonly cited to explain the growth of US malpractice claims.

3.4. The disposition of malpractice claims

Several studies have examined the disposition of malpractice claims, to determine how far the outcome conforms to legal principles and reasons for deviations [Danzon and Lillard (1983), Farber and White (1991), Sloan and Hoerger (1991)]. Claim disposition can be modeled as a process in which the litigants form expectations about the outcome

at verdict (probability and size of award) and expected costs of litigating to verdict. A necessary (but not sufficient) condition for settlement is that the difference between the expected outcomes is less than the sum of the litigation costs. More information accrues as time and the disposition process advance. Variants of this model introduce risk aversion, strategic behavior, etc. An important implication of the simple model is that the disposition process involves nonrandom selection of cases to close at each stage of disposition. In particular, cases that go to verdict are disproportionately those in which the plaintiff or defendant tends to overestimate his chances of winning and in which the stakes are sufficiently large for the difference in expectations to exceed the litigation costs.

These various studies of claims disposition show similar patterns. Of the claims filed with insurance companies, 39.6 percent are closed without filing a legal suit, 53.6 percent are closed after suit but before trial, 1.6 percent are closed during trial, 3.2 percent are closed at verdict and 2.0 percent are closed on appeal. Overall, only 43 percent of claimants receive any payment [Bovbjerg (1995) and sources cited therein]. As the disposition process proceeds, claimants obtain more information about the likely success of their claims, which contributes to the high percentage dropped or settled before verdict. Sloan et al. (1991) show that claims with bad news in the patient's chart from the plaintiff's standpoint are more likely to drop their claims [see also Danzon and Lillard (1983)].

Negligent injuries are more likely than nonnegligent injuries to lead to claims and valid claims are more likely than invalid claims to receive payment [Weiler et al. (1993), White (1994)], but the system is not perfect. Farber and White (1991) examine claim disposition using data on a small sample of institutions with an evaluation of the cause of the injuries by independent reviewers. They conclude that negligence was present in 35 percent of claims, not present in 42 percent, with the remainder uncertain. For claims with negligence, the probability of receiving compensation was 0.66 and the mean payment was $205,000; for claims without negligence, the probability of payment was only 0.16 and the average payment was $41,800. White (1994) reviews data from several studies and concludes that the probability of a claim is 0.026 per negligent injury, 0.01 per non-negligent injury and 0.001 per noninjury. This much higher probability of suit for negligent treatment than for non-negligent treatment should provide a significant deterrent effect, despite the high overall error rate in claiming.

These studies consistently show that the size of awards and settlements is strongly influenced by the plaintiff's economic loss. Jury awards, which are the publicly visible component of the malpractice system, are a tiny but self-selected and atypical subset of claims, including disproportionately claims with high stakes and uncertain merit. Given the selection process in which cases with strong evidence of negligence tend to settle, it is not surprising that doctors win over two thirds of the cases taken to verdict but that awards are very large when the plaintiff prevails. Of cases that receive some payment at verdict or in settlement, the mean award is over $100,000 but the median is under $50,000. The distribution of payments is right skewed (approximately log normal), which partly reflects the underlying distribution of injury severity, with modest

losses for the majority of injuries but a few extremely severe injuries [Danzon and Lillard (1983)]. The largest awards are most likely to be reduced either by the judge after trial or on appeal [Shanley and Peterson (1987)].

These findings suggest that the most extreme criticisms of the tort system as a random lottery are exaggerated. The legal system appears to be quite effective at eliminating invalid claims, paying either zero or relatively small amounts to the majority. Moreover, the apparent shortfall between claim rates and negligent injury rates overall does not necessarily imply that deterrence incentives and compensation are too low or that reforms should be designed to stimulate more claims. Given the existence of professional norms and other regulatory and market-driven quality assurance mechanisms, the optimal deterrence incentive through the tort system is certainly less than without these alternative sanctions, although it cannot be determined precisely. Ignoring deterrence concerns and focusing on the compensation function of tort, compensating small claims through the tort system is probably not cost-effective, given other social and private insurance programs that provide appropriate compensation for minor injuries at lower overhead cost. In the Harvard study, nearly 80 percent of the patients who suffered a negligent injury but did not sue were either fully recovered within 6 months or were over 70 when the injury occurred, suggesting relatively small compensable damages. Although the optimal rate of claims per negligent injury remains an unsettled question, given other deterrence and compensation mechanisms, it is clear that a reduction in number of false positive claims and more speedy elimination of those that do occur would improve the efficiency of the system.

4. Malpractice insurance

The supply of medical malpractice insurance violates several of the predictions of standard insurance theory, two of which are addressed here. First, the structure and pricing of liability insurance contracts makes little use of contractual terms or rating provisions designed to restrain moral hazard on the part of the insured, although in principle the risk is within the control of the insured – if not, the deterrence justification of tort liability must be rejected. In particular, malpractice insurance rarely requires co-payment in the form of deductibles or coinsurance, which are common in first party insurance such as health insurance, and premiums are not generally experience-rated, based on prior claims experience, as is common in automobile liability insurance. Second, although theory predicts that insurance should be available at a price equal to the expected loss plus an expense loading, in fact the price and availability of malpractice insurance has been extremely erratic. The mid-1970s "crisis" was associated with premium increases of over 300 percent in some states and total withdrawal of commercial carriers in other states where regulators denied requested rate increases. The mid-1980s witnessed another so-called "affordability crisis". Availability was less problematic, thanks partly to the supply-side changes that occurred in response to the 1970s crisis. These include the formation of provider-owned mutuals, reciprocals, hospital-owned captive companies and risk retention groups that now write over 50 percent of malpractice insurance

coverage, and joint underwriting associations that, like high risk pools in other lines of insurance, are state-mandated suppliers of last resort. In addition, the risk borne by insurers was reduced by the switch from occurrence coverage, which covers all claims arising out of practice in the policy year, to claims made coverage, which pays all claims filed in the policy year, regardless of the practice year in which the alleged injury occurred, provided that the insured was covered by that insurance company at the time (see Section 4.2).

4.1. Liability insurance contracts: experience rating and co-payments

Shavell (1982) shows that liability insurance need not interfere with the deterrence incentives of liability if premiums are perfectly experience rated, that is, the price of insurance accurately reflects the insured's expected loss. In practice, malpractice insurance has less experience rating than other insurance lines such as automobile liability or workers compensation, where premiums are usually automatically adjusted to reflect adverse claims experience. Malpractice insurance rates are a multiplicative function of limits of coverage (for example, $1 million per occurrence, $3 million total for the policy year); medical specialty; and geographic location. For example, the territorial rate for a base class and basic limits is multiplied by specialty differentials and excess limits factors to obtain rates for other specialties and higher limits of coverage. Rates are generally not based on volume of business, except for a crude part time adjustment, and are not automatically surcharged for claims [Danzon (1985a), Sloan (1991)].

Several studies have confirmed that the distribution of claims, conditional on medical specialty, is highly skewed, with a small number of physicians accounting for a larger number of claims than would be expected if the probability of a claim were uniform and the judicial process entirely random [Rolph (1981), Nye and Hofflander (1987), Sloan, Mergenhagen et al. (1989)]. These findings are consistent with the hypothesis that malpractice claims are disproportionately due to minority of "bad apple" providers. However, at least part of the variation in claims experience could reflect unobserved variation in case mix or volume. Consistent with this interpretation, Sloan et al. (1989) find that board certified physicians and physicians who work longer hours have more claims. The evidence is inconclusive on how far claims are triggered by poor communication with patients as opposed to inferior clinical care. Hickson et al. (1994) find that physicians with a prior history of malpractice claims are more often the subject of subsequent patient dissatisfaction over inadequate time and explanation than physicians who had never been sued. Entman et al. (1994) find that prior claims experience is unrelated to subsequent technical quality of care. Taken together, these findings might be interpreted to suggest that poor communication is a more important determinant of claims than poor clinical care. However, an alternative possible explanation is that deterrence is effective for clinical care, such that physicians with prior claims experience perform better in the future, whereas communication skills are not improved.

The frequency of claims per physician is significantly higher for surgical specialties than for non-surgical specialties, although it seems unlikely that surgeons are consis-

tently more careless than non-surgical specialists. A plausible explanation is that surgical errors are more likely to be severe and causal connections to treatment are more obvious.[10] This suggests that courts distinguish imperfectly between adverse outcomes due to bad luck vs. negligence.

The findings of nonrandom distribution of claims has led some to argue for more experience rating. But Ellis, Gallup and McGuire (1990) show that rating based on Bayesian conditional means with five years of experience would move premiums only modestly towards actuarially fair rates on average, while introducing inequities between physicians with identical underlying risk and exposing physicians to considerable financial risk of inappropriate surcharges. They estimate that under such a rating scheme, a single paid claim would result in a four-fold increase in premiums for most medical specialties. Similarly, Rolph et al. (1991) find that 5 years' prior claims experience has only modest predictive power for future claims experience. Thus, if judicial error is significant, risk aversion would explain the lack of demand for experience-rated policies. Companies that do base rates on prior experience often conduct an independent review, rather than impose an automatic surcharge for all claims [Schwartz and Mendelson (1989)]. The demand for community-rated policies may thus be viewed as insurance against the risk of error by claimants, the courts and the settlement process [Danzon (1985b)].

A related apparent puzzle is the relatively infrequent use of deductibles or other copayments commonly used in other lines of first party insurance to control moral hazard. Several factors may limit the demand for policies with monetary deductibles and copayment. First, being sued entails uninsurable costs of time, in addition to anxiety and threat to reputation; these uninsured costs are probably equivalent to a deductible of several thousand dollars. Second, the potential for a claim in excess of the policy limit implies additional uninsured risk. Third, since liability insurance covers legal defense in addition to indemnity payments, a deductible would reduce the malpractice insurer's incentives to defend claims that could be settled within the limits of a deductible. Fourth, the sorting of physicians into companies acts as a crude form of experience rating, if lower priced insurers with stringent underwriting standards reject physicians with bad prior claims experience. At the limit, physicians who lose their insurance in the standard market may obtain coverage from surplus lines carriers, who charge high premiums, and impose large deductibles and restrictions on coverage [Schwartz and Mendelson (1989)]. Whether these hidden uninsured costs add up to less or more than the socially

[10] Weiler et al. (1993) found significant differences in the adverse event rate across medical treatments-for example, cardiac surgery had three times the adverse event rate (10.8 percent) of general medicine (3.6 percent). However, the differences in proportion of injuries due to negligence were not statistically significant. Surgical complications accounted for 47 percent of all adverse events, but the percent of these attributed to negligence was only 17 percent, compared to 37 percent for non-surgical adverse events. Weiler et al. (p. 53) conclude "Although the total number of negligent injuries inflicted by surgeons was higher than the number of injuries caused by internists, the difference appears to reflect the complexity and riskiness of the procedures performed by the two groups, and thence the very different consequences of momentary lapses".

optimal degree of co-payment is an important but unanswered question. Imposing more co-payment or experience rating by law would increase uninsured risk for physicians and could lead to defensive responses, such as refusal to take high risk patients.

Even if it could be shown that the great majority of physicians take appropriate care most of the time and that most negligent injuries are the result of occasional mistakes by otherwise competent physicians, it certainly would not follow that liability serves no useful deterrence purpose. It is possible that liability has a negligible marginal effect but the average effect is large, such that if liability were eliminated, more physicians would be less careful and make more frequent mistakes. A related argument is that liability is unnecessary because of alternatives such as state professional review mechanisms, hospital quality assurance programs etc. Although in theory these mechanisms could be substitutes, in practice they may be complements and depend on the liability system. For example, casual evidence indicates that hospital and other peer review procedures have been strengthened in direct response to liability. Sloan (1989) report that fewer than 10 percent of physicians with adverse claims experience were disciplined in any manner by professional review boards.

4.2. Premium levels and availability

Malpractice insurance premiums appear to follow a cyclical pattern, with periods of sharp rate increases and limited availability followed by periods of flat or falling premiums. Although over the long term the increase in premiums is explained by the rise in claim costs, the short term cycle in premiums is more erratic than the trend in claims. In the early seventies, premium rates initially lagged behind rising claim costs, necessitating sharp premium increases in the mid-1970s of over 300 percent in some states and withdrawal of traditional insurance carriers in other. In the late 70s, claim costs stabilized and insurance rates fell in real terms. Premiums increased sharply again in the mid 1980s. Similar "crises" occurred in the mid 1970s and 1980s in other "long tailed" lines of liability insurance, such as product liability, municipal and environmental liability, for which claims may be filed many years after the action that allegedly caused the injury. There is concern that the stable rates so far in the 1990s may partly reflect aggressively competitive pricing that may ultimately lead to insurer losses followed by sharp premium increases.

A considerable literature has analyzed the causes of these apparent cycles and crises in long-tailed insurance lines [Cummins and Outreville (1987), Priest (1987), Harrington and Litan (1988), Winter (1988), Harrington and Danzon (1994), Cummins and Danzon (1997)]. Malpractice, product liability and other long-tailed insurance lines have several characteristics that limit the ability of insurers to eliminate uncertainty simply by writing a substantial share of the relevant risk pool. First, the loss distribution per risk is extremely skewed. For malpractice, 5 percent of claims account for over 50 percent of dollars paid [Danzon and Lillard (1983)]. This skewness, combined with the small number of policyholders in a given risk class in many jurisdictions, implies considerable intertemporal variance in the mean loss per insured. Second, the payout tail

on claims may exceed 10 years because the policy covers all claims arising out of injuries that occur in the policy year, but certain long-latent injuries such as cancers may take many years to emerge. Traditionally, the statute of limitations does not begin to run until the injury has, or with due diligence should have been, discovered. The decay of evidence contributes to delay in claim disposition, which may add 5–10 years from claim filing.

Third, although in principle a claim is governed by the rules in effect at the time of the alleged injury, in practice changes in social norms and legal rules may simultaneously affect the loss distribution on all outstanding claims, spanning several policy years and possibly multiple lines of insurance. These common factors cannot be diversified through standard, law-of-large numbers pooling. Changes in social norms and legal rules apparently contributed significantly to the surge in claims for malpractice and other lines in the 1970s and again in the 1980s. Insurance companies failed to anticipate these changes, hence were underreserved and suffered significant shocks to their capital base. The "capacity crunch" theory [Winter (1988), Cummins and Danzon (1997)] posits that shocks in insurance capacity lead to contraction in the supply of insurance, sharp premium increases and possibly lack of availability of coverage for high risk activities or policyholders. However, total withdrawal of commercial insurers, as occurred in some states in the 1970s, is more plausibly explained by denial of requested rate increases by state insurance regulators.

In response to the crises, many malpractice insurers shifted from occurrence coverage, which covers all claims arising out of practice in the policy year, to claims made coverage, which pays for all claims filed in the policy year, regardless of the practice year in which the alleged injury occurred, provided that the insured was covered by that insurance company at the time. Claims made coverage gives insurers greater flexibility to adjust premiums to reflect changes in social and legal norms but, as a result, shifts this risk back to the policyholder. A second response to high premiums or lack of availability of commercial coverage was the formation of provider-owned insurers, including physician-owned mutual and reciprocal companies and hospital-sponsored captive insurers, which now write over half of malpractice insurance premium volume.

The formation of provider-owned insurers was prompted initially by high premiums or withdrawal of commercial carriers. Some were formed with medical society sponsorship to provide insurance to all members; others use selective underwriting to attract better risks. The survival and growth in market share of these mutual carriers, including many that do not practice selective underwriting, suggests that they have persistent advantages relative to stock companies. Provider-owned companies may have certain information advantages over commercial insurers, which facilitates accurate underwriting and premium rating, settlement and merit rating of policy holders. However, commercial insurers can and do involve physician-policyholders in these functions. For example, some commercial policies have been sponsored by state medical societies which play a role in designing the coverage.

Probably the more important advantage of provider-owned insurers is in bearing of the undiversifiable component of risk that derives from socio-legal trends that are com-

mon to all policyholders in the pool. In long-tailed lines of liability insurance, the total risk can be decomposed into the policyholder-specific or idiosyncratic component, which depends on the probability of injury, and the socio-legal or common component, which depends on the resolution of claims, given an injury. Mutuals may have a comparative advantage in bearing the nondiversifiable component [Danzon (1984b), Doherty and Dionne (1993)] and can diversify the nonsystematic component of risk through equity markets, by purchasing reinsurance. For example, mutuals can assess or pay dividends to their members, depending on the realization of the common risk, whereas stock companies must hold capital reserves to protect against such uncertainty.

5. Effects of liability on medical practice

5.1. Formulating the analysis

5.1.1. Theoretical issues

The effect of liability on physicians' behavior is a critical component of any evaluation of the costs and benefits of the malpractice system. The theoretical analysis of physician response to liability depends on several key assumptions. First is the physician's objective function and form of reimbursement. Economic models of tort law generally assume profit maximization or cost minimization as the objective of the potential tortfeasor [Brown (1973), Shavell (1980)]. However models of physician behavior in other contexts generally assume a utility function that includes income and other arguments such as leisure, ethics or reputation or the utility of a representative patient [Pauly (1980), Farley (1986)] to reflect agency concerns. Using such an agency model, Danzon (1994b) simulates physician response to alternative liability rules. The conclusion is that the welfare gains from applying tort liability relative to no liability are much lower if the physician is a reasonably good agent than if he or she is a selfish income maximizer. Welfare gains are also greater under capitation reimbursement than fee-for-service reimbursement, ceteris paribus, because of the incentive effect of capitation to reduce services. (This assumes that the expected benefit of any excess services induced by fee-for-service reimbursement is positive, although less than marginal cost.)

A second theoretical issue is the assumed relation between liability insurance and injury prevention. Ehrlich and Becker (1972) show that insurance and prevention are simultaneously determined responses to injury risk; they may be complements or substitutes, depending on whether the price of insurance adjusts to reflect risk reduction due to prevention. Shavell (1982) demonstrates a similar result in the context of liability insurance. An important implication is that empirical analysis of physicians' prevention response to liability should treat the level of malpractice insurance coverage or premium as a simultaneously determined choice, whereas several studies use the premium as an exogenous measure of liability risk or "climate". A preferred, exogenous measure is the premium rate charged for specified limits of coverage in the physician's locality.

Applying a simple prevention/insurance model to medical liability presupposes a functional relationship between precautions, injuries and a liability loss, and that physicians know this relationship. By contrast, the common view of the legal system presupposes either that the risk of suit is independent of the level of precautions or that it can be influenced by defensive measures such as extra tests, which reduce the probability of suit with little or no effect on the risk of injury. Danzon (1990a, 1990b) models the physician's choice of defensive practices designed to reduce the risk of successful suit as a third alternative to prevention and insurance. Assuming that the physician's objective function includes the patient's utility, in addition to the physician's own income and leisure, this model concludes that extensive defensive measures are unlikely unless physicians place low weight on the patient's utility.

A third issue in analyzing physicians' response to liability is to distinguish between the cost-justified prevention (deterrence) that liability is intended to encourage and wasteful "defensive" responses. This is discussed in Section 5.3.

5.1.2. Empirical issues

Accurate measurement of the effect of liability on medical care requires detailed information on the services provided (M); a comprehensive measure of liability risk (L); and other factors that may affect care, in particular, the patient's insurance and the provider's reimbursement (Z). Measurement of effects on injuries requires, in addition, a comprehensive measure of health outcomes (H). The care response equation can be written:

$$M_{it} = a_1 + a_2 L_{it} + a_3 Z + u_{it}. \tag{5.1}$$

Measurement of all these variables is problematic. If the observed data on medical services, M, are not comprehensive – for example, ambulatory services only – estimated effects may be upward or downward biased, depending on whether the measured medical services are substitutes or complements for omitted services. The liability risk L is multidimensional, but can be proxied by the local price for specified limits of insurance coverage, under the assumption that insurance is accurately rated at the community level; however, the premium paid by the individual doctor confounds price per dollar of coverage (which is exogenous, assuming no experience rating) with level of coverage purchased, which must be treated as endogenous as discussed earlier. The vector of other determinants of provider behavior should ideally control for all other relevant factors. Omission of variables that affect practice and are correlated with L – for example, other locality-specific insurance characteristics such as prevalence of managed care – would lead to biased estimates of a_2.

Estimates of Equation (5.1) using either cross-section or time-series variation, can at best measure the marginal response to changes in liability. However, it is the total, all-or-nothing effect of liability that would be required to evaluate radical policy proposals, for example, for replacing tort liability with a tax-funded, no-fault compensation scheme. Moreover, Equation (5.1) alone cannot distinguish cost-justified deterrence from wasteful defensive medicine in the absence of data on outcomes.

Two types of data have been used to estimate variants of Equation (5.1). The first is surveys undertaken specifically to ask physicians about their response to liability. For example, Reynolds (1987) use AMA survey data that asked about several dimensions of practice that are believed to be most affected by liability, including use of lab test, X-rays, C-sections, referrals, etc. Reynolds et al. conclude that 14 percent of expenditure on physicians' services are a defensive response to liability. Although widely cited and updated [Rubin and Mendelson (1993)], these estimates are subject to several biases. Most obvious, physicians may exaggerate in ascribing their use of costly procedures to liability rather than to other factors, such as financial incentives. Second, the response is fully assigned to defensive medicine, none to cost-justified deterrence. Both of these factors would imply that the Reynolds et al. results are an upper bound on true defensive medicine. Third, because the listed services are only a subset, changes in other substitute or complementary inputs are not accounted for. Fourth, reporting the estimated cost of response as a net increase in expenditure presupposes that this cost is fully passed forward to patients/payers in higher billings, with no incidence on physicians. This assumption may have been valid in the 1980s [Danzon, Pauly et al. (1990)], but is less so with the growth of managed care.

Surveys designed to collect information for purposes other than measuring response to liability are less likely to be subject to reporting bias, but typically contain incomplete information on patterns of practice. The Physician Practice Cost and Income (PPCI) surveys have been used in several studies, but include information on a limited number of ambulatory care price and service characteristics. Reimbursement claims data are more complete but lack detail on other practice characteristics, such as time per encounter.

5.2. *Empirical evidence*

Danzon et al. (1990) and Danzon (1990a, 1990b) analyze several dimensions of physician response using the PPCI surveys of 1976, 1978, 1983 and 1986, which surveyed a different, but nationally representative sample of physicians each year. The liability climate is measured by the premium rate for basic limits of insurance coverage. These surveys span the years of the liability insurance crises but predate the widespread growth in capitation and managed care, which may have changed physicians response and limited their ability to pass on high premium or service costs through additional billings or fee increases.

This analysis concludes that physicians increased their expenditure on insurance less than in proportion to increases in expected liability loss costs. Physicians therefore bore more uninsured risk in states with high liability costs. At that time, liability insurance increases were passed along promptly through higher fees and reimbursement by health insurers. The elasticities of routine office and hospital visit fees with respect to liability insurance rates are between 0.1 and 0.2. This is more than sufficient to pass on the cost of increased expenditure on insurance, assuming no change in volume, since on average physicians spent roughly 4 percent of gross revenues on insurance. Elasticities of reimbursement paid by health insurers are similar to fee elasticities in the 1970s, but

somewhat lower in the 1980s, consistent with increasingly aggressive attempts at expenditure control by insurers. By 1983, the ratio of Medicaid reimbursement to usual fees is negatively related to liability, possibly because Medicaid reimbursement lagged most in urban areas which also tend to have high malpractice costs. This suggests that if, as is often alleged, liability has made physicians less willing to treat Medicaid patients, the relatively tighter constraints on cost pass-through to Medicaid is an important contributing factor. But in general, this pass-through of malpractice costs into higher fees and reimbursement levels was rapid and direct, without requiring an adjustment in physician stocks. The number of physicians per capita, by county, and the rate of change between these years was unrelated to either levels or rates of growth of liability costs.

The excess of the fee elasticities over the level required to fully pass-through the costs of malpractice insurance may reflect several factors: increased physician time per patient encounter, which could reflect improved care; a compensating differential for exposure to uninsured claim costs and/or uninsurable time and non-monetary costs associated with the risk of suit; and a reduction in volume in response to higher fees. There is weak evidence that liability induced physicians to spend more time per patient visit – a possible indicator of more care.[11] By contrast, the frequency of lab tests or procedures was significantly negatively related to liability costs; the frequency of X-rays or fluoroscopies was positively related to malpractice costs in the 1970s but the relationship was negative in 1983. Total number of office visits was negatively related to liability costs, which is consistent with standard constraints on demand and inconsistent with unlimited ability or willingness of physicians to shift demand for defensive purposes, or with demand shifting outward in response to perceived improvement in quality of care or higher expected compensation in the event of injury. This evidence is thus not consistent with significant defensive ordering extra tests, at least in ambulatory care.

On average, physicians' reported net money incomes were not adversely affected by liability costs through 1983, which is consistent with the evidence of a rapid pass through of cost increases to fees and no effects on the geographic distribution of physicians. However, several caveats are in order. First, even if net money incomes were maintained, physicians' real utility may be lower, because of slightly longer hours of work and increased exposure to the uninsured risk monetary and non-monetary costs of being sued. Second, as medical care markets have become more competitive, the ability to pass through premium increases is probably more limited in the 1990s than at the time of these surveys. Third, these estimates of mean effects may obscure significant distributional effects, with losses to some physicians offset by gains to others. A combination of these factors may explain why physicians lobbied for changes in the

[11] A liability-induced increase in time per visit is plausibly consistent with cost-justified deterrence, assuming that fee-for-service reimbursement and a fortiori capitation create incentives for suboptimal physician time per visit in the absence of liability. By contrast, fee-for-service reimbursement is likely to lead to excessive use of tests and procedures, even without liability. If so, an increase in tests and X-rays is more likely to be wasteful defensive medicine.

malpractice system, even if their fees and net money incomes, on average, rose to keep pace with malpractice insurance costs.

Weiler et al. (1993) surveyed physicians to obtain information on their perceived risk of suit and actual suit history, in addition to practice changes. These data permit estimates of whether self-reported practice changes are statistically related to self-reported perception of liability. They find that physicians with prior claims were significantly more likely to explain risks to patients, and that those with high perceived risk of suit were significantly more likely to order more tests or procedures and reduce the number of patients or procedures (such as GPs dropping minor surgery). This study also found that physicians incurred significant financial and nonfinancial costs of being sued, confirming anecdotal evidence that physicians perceive a significant tort threat despite extensive liability insurance.

5.3. Defensive medicine

Defensive medicine may be defined as liability-induced changes in medical practice that entail costs in excess of benefits, so would presumably not be desired by an informed patient, given their insurance coverage, and that are intended to reduce the physician's risk of being sued. Defensive medicine is thus a form of supplier-induced demand [SID, see McGuire (2000)]. Distinguishing empirically between liability-induced defensive medicine and insurance-induced waste (use of medical services with expected benefits less than social cost because the patient is insured) is problematic, because insured patients have little incentive to refuse any treatment with non-negative benefit. Physicians may face little patient resistance to low benefit services that increase their income (under fee for service) and/or reduce their risk of suit. However, such practices and the pass-through of associated costs to payers should decline with the growth of managed care and capitation, to the extent that capitation internalizes the cost of defensive practices to medical providers.

Several studies have focused on the effect of liability on the use of certain procedures that are allegedly commonly used for defensive purposes. Using patient level data from 31 hospitals in New York in 1984, controlling for patient and physician risk factors, Localio et al. (1993) find that use of caesarean section is significantly positively related to three measures of malpractice risk: premium rates in the locality; the mean perceived risk of suit in the locality, as reported in surveys of physicians; and the prior claim history of the hospital and the medical staff in which the physician practices.

In order to control for possible bias due to unobserved omitted variables that is present in the cross-section studies discussed so far, Kessler and McClellan (1996) apply a difference-in-differences estimator to data on Medicare patients hospitalized for treatment of acute myocardial infarction (AMI) or ischemic heart disease (IHD) over the period 1984–1990. They use the enactment of tort reforms as a measure of change in the state-specific liability climate. They estimate the effect of tort reforms on total hospital expenditures in the year after the AMI (a measure of the intensity of treatment) and on two measures of outcome, mortality or repeat hospitalization for AMI or heart disease within one year of the initial illness. A key assumption is that statutory changes in

tort law affect providers' perceptions of liability risk and hence affect practice patterns. The expenditures and outcomes equations control for patient demographic characteristics, state legal and political characteristics, and state and time fixed effects. The effects of tort reforms are estimated as the difference in time trends between states that changed and states that did not change their laws.

Kessler and McClellan find that tort reforms have a significant negative effect on cost but no significant effect on outcomes, which they interpret as evidence that higher liability creates incentives for socially excessive care or defensive medicine. This is an imaginative and careful study, but the possibility remains that the estimated effects may reflect some unobserved, correlated factor. In particular, it is plausible that states with aggressive managed care were more likely to adopt tort reform, since managed care limits the ability of providers to pass through liability-related costs. If so, the cost savings attributable to tort reform could in fact be due managed care. Consistent with this, in preliminary findings from work in progress Kessler and McClellan find that controlling for managed care penetration reduces but does not eliminate the negative association between tort reform and expenditures.

6. Overall evaluation of the malpractice system

A full evaluation of the medical malpractice system must weigh the costs against the benefits. Although the available data are insufficient for a definitive analysis, rough calculations are informative.

6.1. Costs of administration and defensive medicine

Malpractice insurance premiums are less than 1% of total health care spending; nevertheless, this small percentage is roughly $10 billion [Rubin (1993)]. From the patient's perspective, if tort liability is solely a form of compulsory insurance system for iatrogenic injuries, for which premiums are included in the cost of health care, then it is grossly inefficient relative to alternatives. Roughly 40 cents of the malpractice premium dollar reaches the patient as compensation. Of the remainder, 40 cents per $1 premium is spent on litigation, roughly equally divided between plaintiff and defense attorneys and 20 cents is insurance overhead [Danzon (1985a)]. Compensation through tort liability thus carries a loading charge of $1.50 per $1.00 of compensation, compared to less than 10 cents per $1.00 of compensation for large group first party insurance. This overhead cost is pure waste if and only if the investigation into cause and fault has no deterrence benefit. Tort also entails additional real but hidden costs, including the time and anxiety costs of litigation; a relatively long mean delay (several years on average) from the occurrence of an injury to receipt of payment; and uncertainty as to timing and amount of compensation, whereas insurance is intended to eliminate variance in income.

The second potential cost of the malpractice system is defensive medicine. As discussed above, if courts lack good information about the optimal standard of care, physicians may anticipate that they can reduce their probability of being found liable by taking highly visible but unnecessary precautions, such as ordering tests beyond the level that is desired by patients given their insurance coverage, with expected benefits less than cost. Defensive medicine remains an unmeasured deadweight loss of the liability system. Theory suggests that it is likely to be larger under traditional fee-for-service reimbursement than managed care.

Given these excess overhead costs of insurance and defensive medicine, the clear conclusion is that if the sole function of liability is to provide compensation, this can be done more efficiently through other private and social insurance mechanisms. The critical question is whether these excess costs are matched by at least equivalent deterrence benefits, such that overall the benefits outweigh the costs.

6.2. Deterrence benefits

Measuring the deterrence benefits of tort liability requires estimating its effect on the incidence of negligent injuries, which poses severe measurement and estimation issues. The econometric challenge is somewhat easier for automobile accidents where the frequency of accidents is more readily observable and changes in liability rules across states and over time provide clear and measurable differences in the liability regime – for example, change from third party negligence rules to first party no-fault rules have occurred in several states in the US and in Quebec in Canada. The evidence from these studies, summarized in Sloan (1998), confirms that tort liability is a deterrent to unsafe driving, mediated in part by associated changes in the availability and price of insurance.

The only credible study of deterrence of medical negligence is from Weiler et al. (1993). This study estimated the relation between proportion of negligent injuries and claims per negligent injury across 49 hospitals in New York state.[12] An instrumental variables approach is used, using urbanization and population density as instruments, to control for potential bias due to the endogeneity of negligent injuries in the denominator of the deterrence measure, claims per negligent injury. The point estimate is negative but not statistically significant. Taking this point estimate at face value and extrapolating would imply that tort liability reduced the rate of negligent injuries per admission by 29 percent (from 1.25 with no liability to 0.89 with the current system) and reduced the overall rate of medical injuries per admission by 11 percent (from 3.7 to 3.3). The failure to find significant effects may be influenced by the small sample size (49 hospitals) and imperfect instruments available. Moreover, at best these data would estimate the marginal effect of changes in liability over the limited range of variation in the New

[12] Using the proportion of negligent injuries as the dependent variable controls for unobserved case mix across hospitals which might lead to variation in rates of nonnegligent injuries

York sample. Considering these intrinsic limitations that bias against finding significant effects, together with other evidence that physicians do perceive a significant risk of suit and change their behavior in response to liability, Weiler et al. conclude that liability plausibly does have a significant deterrent effect.

This empirical evidence on deterrence benefits is consistent with rough calculations by Danzon (1985a), that under quite generous assumptions about the costs of defensive medicine, the malpractice system would pay for itself (yield positive net benefits) if it reduced negligent injury rates by at least 20 percent, ignoring such intangible benefits as retribution or fairness. As Weiler et al. (1993) note, the point estimate from the Harvard study, that negligent injuries are reduced by 29 percent, easily passes this test. This perhaps surprising conclusion is possible, despite the high overhead expense of the malpractice system, because of the low rate of claims per negligent injury. Since the high administrative loading is incurred only on the small fraction of injuries that lead to a claim, a modest percentage reduction in injury rates is sufficient to offset reasonable estimates of overhead and defensive medicine costs. Even if the benefits of the current system do outweigh its costs, however, the search for marginal improvements or more cost-effective alternatives remains an important policy question.

7. Traditional tort "reform"

Since the mid-1970s, most states have adopted one or more changes in their traditional tort rules for medical malpractice. Follow standard usage, these changes are referred to here as tort "reforms", without implying any endorsement of such changes as desirable.

The economic analysis has shown that the most extreme criticisms of the system as a costly lottery are unfounded: negligent injuries are more likely to result in a claim and compensation than non-negligent injuries and awards are significantly related to the magnitude of loss. The various pieces of evidence suggest that medical practice patterns are affected and may plausibly provide a sufficient deterrent effect to outweigh the costs. Thus it is plausible but not proven that the system overall yields a positive net social benefit. Nevertheless, any reform that reduces the deadweight costs of litigation and defensive medicine or improves the efficiency of deterrence or compensation without increasing litigation or overhead costs would improve the efficiency of the system.

The economic criterion for evaluating a proposed reform is thus, Is it likely to reduce the deadweight loss of litigation and defensive medicine, or improve the efficiency of deterrence and compensation, recognizing that the practical choice is between imperfect alternatives? By contrast, most actual reform proposals aim primarily to reduce measurable claim costs and liability insurance premiums or budgetary costs to health care providers. This budget focus is likely to result, at best, in simply shifting costs from medical providers to patients and taxpayers; at worst, total social costs may actually increase if, for example, deterrence incentives are weakened.

7.1. The structure of awards

7.1.1. Limits on awards for nonmonetary loss

Roughly half the states have enacted caps on awards; most limit only non-monetary loss but a few limit the total award. Award caps have been estimated to reduce mean payout per claim by up to 40 percent [Danzon (1984a), Zuckerman, Bovbjerg et al. (1990), Harrington and Danzon (1994)] and premiums by somewhat less. Such large effects are possible, although the caps directly constrain only a small percentage of cases, because roughly 5 percent of cases account for 50 percent of dollars paid.

Economic evaluation of tort awards concludes that, while some limits are desirable, single caps are at best a second best approach. A tort award in principle serves two functions: it provides compensation to the plaintiff and imposes a fine on the negligent defendant, assuming no liability insurance. The traditional guideline for tort awards is to "make the plaintiff whole". This full compensation principle is unlikely to be optimal for either compensation or deterrence, at least in the case of seriously disabling injuries that result in an "irreplaceable loss" [Cook and Graham (1977), Spence (1977), Danzon (1984b)]. Since tort compensation is a form of compulsory insurance that is tied to the purchase of medical care, optimal compensation is the amount that consumers would choose to purchase voluntarily, given the expense loading of the physician's liability insurance (see Appendix). Insurance can only transfer money from the healthy to the disabled state, but money is an imperfect substitute for an irreplaceable faculty or possession, and transferring funds is costly. Optimal insurance with zero load equalizes the marginal utility of income in the injury and no-injury states. With a state-dependent utility function, optimal compensation for an irreplaceable loss could be more or less than full compensation, depending on whether the injury raises or lowers the marginal utility of income. With the high loading on malpractice insurance, optimal compensation is presumably lower.

If victims incur an uncompensated loss after optimal compensation, then optimal deterrence may require that the defendant pay a fine in addition to the compensatory award. This optimal deterrence fine depends on consumers' willingness to pay for risk reduction, given optimal compensation. To illustrate, a bachelor with no heirs may be willing to pay large sums to reduce his risk of death even though he might choose not to buy life insurance. More generally, the optimal deterrent fine, conditional on optimal compensation, is inversely related to the extent to which market prices for medical care internalize patients' willingness to pay for risk reduction and to the defendant's uninsurable costs of suit, such as time and reputation [Spence (1977), Danzon (1985b) and Appendix]. The fine should be paid to the state and refunded as a subsidy to the risky activity, in order to preserve appropriate relative prices

Traditional tort rules also permit a two part damage award, consisting of a compensatory award for monetary and nonmonetary loss and a punitive award, in cases of recklessness, wanton or willful misconduct. However, the two part system in practice differs from the theoretical ideal in several ways. Punitive awards are based on the defendant's conduct, whereas ideally they should reflect consumers' willingness to pay for

prevention; punitive awards are paid as additional compensation to victims, rather than to the state; and compensatory awards aim to provide full compensation of monetary and nonmonetary loss, regardless of consumers' willingness to pay for insurance.

Theoretical analysis alone cannot determine optimal compensation for an irreplaceable loss, because the marginal utility of income is unobservable. However, the empirical evidence that consumers do not voluntary buy coverage for noneconomic loss in any other private or social insurance program suggests that such coverage may not be worth its cost. The lack of a voluntary market for insurance of nonmonetary losses may reflect severe ex post moral hazard of exaggeration of such losses which cannot be objectively measured. Assuming that this moral hazard of loss exaggeration is at least as severe in the tort system, the evidence from private choices supports the case for limits on compensation for nonmonetary loss through the tort system.

Several studies have therefore concluded that limits on awards for nonmonetary loss would improve efficiency of compensation in the US tort system [Danzon (1984b), Bovbjerg, Sloan et al. (1989)]. Many European countries already have such limits. The preferred approach is a schedule based on the severity of injury and the plaintiff's life expectancy, in order to approximate the ideal of equalization of the marginal utility of income. By contrast, in states that have enacted limits on awards since 1975, the great majority have adopted single caps on nonmonetary loss for all cases, say $500,000, rather than scheduled benefits for nonmonetary loss. This may be too low for young, severely injured plaintiffs, excessive for older patients or minor injuries.[13] This preference for single caps may reflect concern to avoid "injury severity creep" or litigation over whether the differentials are fair – issues that are ignored in the theoretical analysis of scheduled benefits.

In addition to providing more optimal compensation, scheduled awards are also expected to reduce litigation expense, by reducing the marginal payoff to investment in litigation effort. Limits on compensation for nonmonetary loss are unlikely to undermine deterrence, because very high awards are typically not used for rating individual (as opposed to class) liability premiums, being viewed as random bad luck. If deterrence derives primarily from the uninsured time, anxiety and reputational costs, these are reportedly invariant to the outcome of the claim [Weiler et al. (1993)]. Thus scheduled limits on awards could improve the efficiency of compensation and reduce litigation expense, with no effect on deterrence.

[13] There may be a further objection to single caps, if actual awards already tend to undercompensate for the economic loss of severely injured patients more than for minor injuries (Sloan, Githens, et al. 1991). The evidence on this point is inconclusive. In principle, one would like to compare the patient's compensation at verdict, net of attorney fees, to the economic loss incurred. In practice, the sample of cases closed at verdict is too small to permit such estimates. Inference from the much larger sample of out-of-court settlements is problematic because settlements reflect the expected verdict, discounted to reflect the plaintiff's probability of winning, net of the differential in litigation costs (see Equation 1). If the plaintiff's probability of winning is inversely related to economic loss in closed claims data, because of fixed costs of going to court (see Danzon and Lillard, 1983), this could account for an inverse relation between the compensation/economic loss ratio and injury severity.

7.1.2. Periodic payments

Under traditional tort rules, compensation for future damages is paid as a lump sum equal to the discounted present value of future payments. Several states now permit periodic payment of compensation for future damages. The intended level of patient compensation may often be provided at lower cost to the defendant through purchase of an annuity or other financial instrument, if courts tend to be more conservative than financial markets in estimating interest rates, inflation and life expectancies. The amount of such future payments should be fixed at the time of claim disposition. Periodic payments that are contingent on the actual reported loss provide more than optimal insurance and tend to undermine incentives for rehabilitation [see Rea (1981) for theoretical analysis; Butler (1983) for empirical evidence from workers compensation].

7.1.3. Collateral source offset

Under the traditional collateral source rule, tort awards in the US are not reduced by the amount of compensation that the patient receives from private or public insurance. Such offset occurs automatically in countries such as Sweden, the UK and Canada, where medical costs that are covered under national health systems are not compensable in tort. Since 1975, many states in the US have provided for offset from the tort award of certain forms of insurance, to avoid a windfall of double compensation to the plaintiff and reduce malpractice premiums.

With perfect information and costless transacting, the collateral source rule would be irrelevant because consumers could contract around it [Coase (1960)]. Consumers could choose first party insurance that makes no payment in the event of a tort award or with subrogation, whereby the first party insurer assumes the plaintiff's tort claim for covered expenses. However, with nonzero information and contracting costs, the collateral source rule matters. Which rule is on balance more efficient is an empirical question. Subrogation preserves the full internalization of injury costs to the tort defendant and hence preserves stronger incentives for deterrence. By contrast, collateral source offset undermines deterrence by shifting costs from the tort defendant to other insurance programs and by reducing the plaintiff's incentive to bring a claim because of the lower expected award. Empirical evidence confirms that collateral source offset rules have not only reduced claim severity but also claim frequency, consistent with the prediction that lower awards reduce the incentive to file [Danzon (1984a, 1986)]. However, because subrogation may entail higher transactions costs than collateral source offset, the optimal mechanism for eliminating double compensation remains an unresolved empirical question.

7.2. Reducing litigation costs

Litigation expense is at least partly a voluntary investment made by the litigants, given the costs and expected payoff. A simple model of rational investment in litigation yields

important implications for reform. First, measures that reduce the elasticity of awards with respect to litigation effort, such as damage caps or scheduled benefits, should reduce litigation investment. Second, measures that reduce the cost per unit of litigation input, such as substituting arbitration for more costly court proceedings, may increase the number of claims filed and total litigation expense per case could rise or fall; efficiency effects are uncertain, but an increase in outlays is certainly contrary to the intent of such reforms. Third, measures that reduce litigation inputs will also affect outcomes, hence a full evaluation must consider effects on compensation and deterrence. In particular, the optimal amount of litigation depends on the social benefits of injury deterrence and the private benefits of compensation, and the divergence between private and social costs of litigation that results because each litigant imposes costs on the opposing party and on public financing of the courts. Shavell (1997) analyses the implications for liability reform of this divergence between private and social benefits and costs of suit.

7.2.1. Limits on contingent fees

In the US, plaintiff attorneys on medical malpractice and other personal injury litigation are typically paid on a contingent basis, that is, they receive a fee if and only if they win the case. The most common fee is one third of the award or settlement, with a range of 25–50 percent. Investing in litigation with uncertain payoff is a risky business. An attorney with a portfolio of cases can more efficiently bear this risk than an individual plaintiff, for whom the legal expense if paid as an hourly fee may be a significant fraction of wealth. Contingent fees therefore provide a potentially more efficient allocation of this risk than hourly fees. Nevertheless, contingent fees have traditionally been banned in the UK and Canada.

In the US, several states have adopted sliding scale limits on contingent fees. The allegation is that contingent fees stimulate an excessive number of suits and an excessive willingness to reject settlements and gamble for large jury verdicts. Theoretical analysis predicts that number of claims filed would be higher with a contingent fee, but appropriately so, because risk aversion would deter many plaintiffs from filing valid claims with an hourly fee. More generally, the effect of contingent vs. hourly fees depends on risk preferences and on competition and information in the market for legal services [Danzon (1983)]. If attorneys compete for cases based on the fee percentage and the expected award, then contingent fees may induce a private first best optimum, whereas risk averse plaintiffs may bring too few cases and invest suboptimally per case if required to pay an hourly fee regardless of the outcome. This disincentive is higher under the English rule that the loser pays all costs; however, it may be mitigated in practice by legal aid (see Section 10).

The objective of limits on contingent fees is unclear and effects of such limits on claim frequency and disposition – and a fortiori on efficiency – are uncertain. If the objective is to reduce large awards, this could be achieved more accurately by direct limits in the form of scheduled benefits. If the goal is reduce litigation expense, mea-

sures to reduce uncertainty and the ability of litigants to influence the outcome would deter investment by both sides.

If the goal is to reduce frivolous suits, a cost-shifting English rule, that assigns both sides' legal costs to the losing party, is a more promising approach. There is a concern that this would eliminate many valid claims if plaintiffs are risk-averse; on the other hand, if plaintiffs are judgment-proof or if defendants choose not to enforce the rule, there would be little effect. An alternative is to apply the rule to the plaintiff's attorney, if paid on a contingent basis, rather than the individual plaintiff. Such a rule would increase the plaintiff attorney's incentive to reject cases with weak evidence of negligence. Applying the English rule to the plaintiff's attorney would almost certainly lead to an increase in the equilibrium contingent fee percentage, in order to compensate plaintiff attorneys for the added risk of paying the other side's costs if they lose.

7.2.2. Alternative dispute resolution (ADR)

Several states have adopted forms of alternative dispute resolution (ADR) that are intended to eliminate frivolous claims, expedite claim resolution and reduce litigation expense. Screening, mediation panels and nonbinding arbitration use less formal rules but operate within the traditional court system, whereas binding arbitration replaces the judge and jury with an arbitration panel selected by the litigants.

The effects of these procedures depends on their effect on the incentives and constraints of the litigants. Theory and evidence indicate that mandatory screening, without significant penalties for appeal and without the panel's findings being admissible evidence in court, may simply add an additional tier of delay and costs. For ADR to reduce litigation delay and costs, it must create incentives for the parties to substitute the informal process for more costly trial in a large percentage of cases. This implies that formal arbitration proceedings should be binding. For less formal procedures, the parties should face significant penalties for proceeding to trial against the recommendation of the panel. For example, the early neutral evaluation (ENE) program that has been adopted in northern California provides each side with information about the other's case, through prompt and neutral evaluation [Rosenberg and Folberg (1994)]. If combined with a system of early binding offers and a fee-shifting rule for frivolous rejection of an offer and continued litigation, the costs and delay of claim disposition could in theory be significantly reduced. An early binding offer system, combined with the English rule, creates incentives for each party to act on their true information, whereas bluff and strategic manipulation are penalized. By contrast, screening and mediation, without significant penalties for strategic post-screening behaviour, simply increase delay and costs.

7.3. Quality and standards of care

Many states have enacted measures intended to encourage quality of care through peer review, practice guidelines etc. The federal government has established the National

Practitioner Data Bank, to which insurers must report all claims paid on behalf of practitioners, and hospitals and states must report significant disciplinary actions. By requiring hospitals and other institutions to check a physician's prior experience with the databank before making a staff appointment, the intent is to prevent miscreant doctors with a bad record in one state from simply moving to another state. Whether the benefits of this system outweigh the costs of data collection and risk of misuse of the data by unauthorized parties remains an unanswered question.

The proliferation of practice guidelines may, in theory, simultaneously serve to improve care and reduce liability errors. Medical guidelines are promulgated by approved bodies to provide guidance to physicians on best practice, hence may reduce the incidence of negligent injury. In addition, they may serve as a defense against malpractice claims. Maine is undertaking an experiment in which adherence to promulgated guidelines is a full defense to a malpractice claim [GAO (1994)]. The net effect of such an approach depends on the optimality of the guidelines, whether they can be used for both defensive and offensive purposes, and whether a significant fraction of care can be routinized in this fashion.

7.4. Enterprise liability

Enterprise liability would shift the locus of liability from the individual physician to an enterprise such as a hospital, HMO or health plan. Proponents of hospital-based enterprise liability [Weiler (1991), Abraham and Weiler (1994)] argue that it would improve deterrence and reduce litigation costs. Improvement in deterrence could occur if hospitals have better information than individual doctors and the authority necessary to implement systems-based loss control and quality assurance programs. Moreover, the incentive to adopt such measures might be strengthened if the larger risk pool increases actuarial credibility and hence permits more accurate experience rating of malpractice premiums at the enterprise level than is possible at the individual doctor level. Reduction in litigation cost could occur because the enterprise would be the sole defendant, whereas currently it is common to sue multiple doctors as well as the hospital, and each of these defendants may hire separate counsel.

The arguments against replacing individual doctor liability with enterprise liability of the hospital are several. First, hospitals already have strong incentives to take those precautions that are within their control, including monitoring of staff, since plaintiffs already name the hospital as a co-defendant if there is any possibility of involvement. Moreover, hospitals already frequently arrange for – and sometimes provide through a captive – the liability insurance for members of their medical staff. It is not clear that enterprise liability would add significantly to the existing information or incentives for system-wide loss prevention measures. On the other hand, the deterrence of individual doctors would be weakened unless hospitals implement increased surveillance measures sufficient to offset the elimination of the individual deterrence incentive on doctors.

Second, the savings in litigation costs could be small. Individual physicians would presumably still be required to testify in order to determine what actually occurred. As

long as the liability rule is a negligence rule, showing negligence would require showing that some member of staff failed to take appropriate precautions, and this would require individual testimony, for which physicians might continue to retain their own counsel. Third, for physicians who have affiliations with multiple hospitals but also some ambulatory practice, there would be either duplication or ambiguity of coverage for their ambulatory practice. Fourth, with the decline in the importance of hospitals in the delivery of care and growth in other institutional arrangement, including integrated systems, large physician group practices, etc., it is increasingly anachronistic to view the hospital as the focus of care and hence as the best locus of liability.

These issues can only be resolved empirically. If enterprise liability is potentially efficient, it could already be adopted by voluntary contract between hospitals and their medical staff. In fact, such contractual enterprise liability is already the norm in at least one staff model HMO, in most teaching hospitals and in other contexts where physicians are salaried hospital employees. However, it has not occurred widely between hospitals and physicians who are independent contractors or in the looser, increasingly common network and independent practice HMOs. Plausibly, in such network environments, the hospital or HMO has neither the information nor the authority to control the practice of individual providers, hence retaining individual liability is more efficient. As noted, in such arrangements the hospital, integrated system or HMO may arrange for the purchase of insurance by participating physicians, in part because of their common interest in loss control.

The liability insurance products offered on the market are adapting to meet the needs and risks of these new institutional arrangements. As long as the market for health care requires providers to compete on cost and quality, providers have incentives to contract for assignments of liability and insurance that offer the best trade-off in terms of cost and deterrence incentives. This may include several variants of total or partial enterprise liability, depending on other institutional factors. In this environment, enterprise liability that is assumed voluntarily is to be welcomed. By contrast, a uniform, mandatory requirement could distort the natural evolution of the delivery system, distorting deterrence and possibly care delivery, with no evident benefits.

8. Radical alternatives

8.1. No-fault programs for iatrogenic injuries

No-fault programs provide compensation for injuries caused by medical care, without regard to the fault or negligence of the medical provider. Some proposals would shift from a negligence rule to strict liability, and shift the locus of liability from the individual physician to an enterprise such as a hospital or health plan [Weiler (1991), Weiler et al. (1993)]. An analogy is drawn with the workers' compensation program, in which employers are strictly liable for work-related injuries, without regard to fault. Other variants would compensate all iatrogenic injuries from broad-based taxes rather than premiums paid by medical providers. In addition, most no fault proposals provide for claims adjudication through an administrative agency rather than the courts; benefits

are usually limited to monetary loss, with collateral source offset and at most modest scheduled payments for pain and suffering. The intent of all these proposals is to reduce litigation delay and expense and to provide compensation to more victims of iatrogenic injury. Virginia and Florida have established no-fault programs of compensation for severe, birth-related neurological injuries caused by medical care, with some general funding.

An evaluation of the efficiency effects of a broad-based no-fault scheme for medical injuries must consider effects on all costs, including the costs of injuries, prevention and overhead. Since most of the proposed programs change several components simultaneously, it is important to identify the marginal effects of the individual components of a proposed program, distinguishing features that could be adopted within the current negligence rule and features that are intrinsic to no-fault programs.

Proponents claim three sources of savings from no-fault programs, relative to the status quo. First, cost per case would be lower due to lower payments of nonmonetary loss and collateral source offset. Both of these changes could be adopted without changing the fault-based rule of liability and both – particularly collateral source offset – simply shift rather than reduce social costs. Second, no fault is usually combined with enterprise liability, with associated advantages and disadvantages discussed earlier.

Third, it is said that eliminating negligence as a condition for compensation would reduce litigation expense. Proponents point to the lower overhead expense ratio of the workers' compensation system in the US or the quasi no-fault systems of accident compensation in Sweden and New Zealand. However, none of these systems provides a good analogy for a no-fault system for medical malpractice, because of differences in context and structure.

The argument that no-fault would reduce litigation costs rests on the assumption that it would be simpler, less litigious and less costly to define a compensable event as a medical injury rather than a negligent medical injury. However, evidence from workers' compensation is not necessarily persuasive because of the difference in context: workers are generally in good health, hence the occurrence of a work-related injury is easy to define. This demarcation is less clear and litigation costs are correspondingly higher for occupational diseases and cumulative trauma than for acute injuries. For iatrogenic injuries, whether an imperfect medical outcome is an iatrogenic injury rather than an imperfect cure within the range of normal risk requires an implicit assumption about appropriate care and the probability distribution of outcomes with and without appropriate care. Since an operational definition of medical causation presupposes a standard of appropriate care, it would often be a minor additional step to determine whether that standard had in fact been met. Making a related point, Epstein (1978) argues that showing cause would often require showing some "defect" in treatment, which is very similar to showing negligence. Weiler et al. (1993) recognize that "in cases of medical omission, the judgement about whether a patient's disability was caused by medical management actually rests on an implicit identification of fault on the part of some provider". Since most injuries can be framed as failure to take some precaution that would have reduced the risk of adverse outcome, this implicit equivalence of a causation test and a fault test

potentially applies far more broadly than the set of injuries that might be classified as errors of omission under the current system.

The medical reviewers in the Harvard study were able to make more reliable judgments about causation (adverse events) than about negligent events [Weiler et al. (1993)], which is one factor leading Weiler (1991) to conclude that litigation would be less costly under strict liability rule than under a negligence rule.[14] Such consistency might be considerably less with lay adjudicators in the adversarial context of actual litigation than in the clinical environment of the study.

The Swedish and New Zealand systems, also cited as evidence for the low overhead costs that could be realized by no-fault systems, also provide a misleading analogy. Although both eliminate the terminology of fault or negligence, medical causation is a necessary but not a sufficient condition for compensation [see Section 10, and Danzon (1994a, 1994d)]. The Swedish system retains a notion of medical error. More important, both Sweden and New Zealand owe their low litigation percentages partly to the fact that compensation is their sole function, with no attempt at deterrence. Since providers are not financially liable or exposed to sanction through these systems (with minor exceptions noted later), providers have no reason to oppose – and some reason to support – compensation for their patients. This is very different from no fault with strict provider liability for costs proposed in the US. Moreover, if compensation is denied, patients have much more limited right of appeal than in the US tort system. Thus the low overhead costs reflect lack of incentives for either party to contest the administrative decision on compensation. Whether total social costs of iatrogenic injury are higher, due to increased frequency of iatrogenic injury that offsets any savings in litigation expense remains an important but unanswered question.

Several models for financing no-fault programs have been proposed for the US, with different expected effects on overhead costs and deterrence. The first – and least likely – is to impose strict liability on individual physicians with experience-rated premiums. This would expose individual physicians to unacceptable financial risk, given their relatively small patient load, hence would be inefficient for risk pooling and entail high uninsurable time costs on physicians of defending the much larger number of claims. Recall that if all iatrogenic injuries are compensable, regardless of fault, at least a tenfold increase in the number of claims might be anticipated, based on the New York and California data, even under the optimistic assumption of no invalid claims.[15] If courts err in dismissing false positive claims, providers would have strong incentives to avoid the sickest patients. Strict liability with collateral source offset would create incentives

[14] The high degree of reliability may partly reflect the fact that the medical reviewers were highly trained and used an elaborate Adverse Event Form, which structured their decisionmaking. Reliability might be considerably less if decisions are resolved through an adversarial process and with lay adjudicators. Even the random sample of physicians surveyed showed "marked variation ... in their willingness to label certain outcomes as iatrogenic" [Weiler et al. (1993, p. 125)].

[15] The number of claims could be reduced by excluding minor injuries [Weiler (1991)], but this shifts the cost to the individual or other insurance programs, without reducing costs.

for providers to avoid uninsured patients, for whom the expected liability cost, conditional on an injury, would be higher than for an insured patient.

The second alternative is to place no-fault (strict) enterprise liability on hospitals or health plans [Weiler (1991), Abraham (1994)]. The case for and against hospital-based enterprise liability has been discussed. As noted, if enterprise liability is potentially efficient, it can already be adopted by voluntary contract between hospitals and their staffs, regardless of whether the liability rule is strict liability or negligence. In fact, such contractual enterprise liability has only been adopted in limited circumstances, usually where doctors are in a close and exclusive relationship to a single hospital, as in a staff model HMO or teaching hospitals.

The third financing alternative for a no-fault program is a broad-based tax on medical providers, insurance companies, or general revenues, as in Virginia and Florida. Financing by a tax on medical providers eliminates individual deterrence, but retains the internalization of costs to the health care industry (assuming accurate adjudication of claims). Financing from general revenue taxation eliminates all internalization of costs and deterrence, hence is pure social insurance. It is arguably neither efficient nor equitable to single out victims of medical injury for special compensation, unless there is a deterrent benefit. Compensation can be provided more cheaply through broad-based private insurance and social insurance programs, such as Social Security Disability, Medicare and Medicaid, which provide compensation without regard to cause. Incurring the cost of determining that a particular condition was caused by medical care, rather than genetic or other factors, is worthwhile only if this information is used to promote deterrence.

In theory, deterrence could be preserved despite broad-based funding if the program itself brought suits for negligence against medical providers. Such decoupling of compensation and deterrence could arguably provide prompt compensation for medical injuries, regardless of fault, while preserving deterrence. However, in practice this decoupling approach could increase overhead costs because two actions would be required, one on causation and one on negligence. Moreover, significant tax financing would still be necessary to pay for the nonnegligent injuries. Such compensation is hard to justify on equity grounds when persons in similar condition from other causes, for example birth defects, would not be eligible. Excess burdens of tax financing would add to the real social costs.

8.2. Private contracting

Tort liability establishes a form of mandatory compensation tied to medical care that may provide rules of liability, compensation and dispute resolution quite different from those that patients would prefer if given the choice. Since medical injuries occur in a context in which the parties are in a contractual relation, there is a prima facie case for permitting the parties to contract out of judicially mandated tort rules [Epstein (1978), Havighurst (1995)]. Such contracts might specify the circumstances for liability (for example, gross negligence only), the rules of damages (for example, economic loss only), and the rules and forum for dispute resolution (for example, arbitration).

One objection to private contracting is that patients are poorly informed and are in no condition to consider such issues when they are in need of medical care – indeed this is the basic rationale for exposing medical providers to tort liability. Consistent with this view, courts have generally overturned contracts entered into at the point of care – for example, a contract providing for arbitration signed when the patient was admitted to a hospital. However, contracts entered into as part of the health insurance agreement are not signed under duress and have generally been upheld by the courts.

Although contractual reassignments of liability are rare to date, this may reflect the difficulty of internalizing benefits to those signing the contract than to lack of potential interest on the part of patients and providers. Under traditional fee-for-service insurance with free choice of provider, patients have little incentive to adopt contracts that limit their tort rights since they cannot realize the full savings from lower premiums unless all the doctors and patients in the area adopt the same contract. At most, patients might face a lower co-payment if they chose a provider who had signed a contract with limited tort rights and hence had lower fees, but the saving would be only a fraction of the total. Moreover, if the provider tried to target the savings through lower fees to the patients who had agreed to the liability restriction, this might be viewed as a contract of adhesion by the courts because it would be patient-specific.

By contrast, if managed care plans can lock in patients to the providers that have adopted a cost-reducing contractual change, then the full savings can be passed on to the patients through lower premiums for the insurance, which avoids the legal problem of contracts at the point of service. Thus managed care offers the potential for more contractual specificity both with regard to the conditions of compensation for iatrogenic injury and the coverage of medical care (see Section 9.1). Stipulating such provisions as part of the health insurance contract would permit consumers to make informed choices before they need care, which in turn increases the likelihood that courts would uphold the contracts.

9. Liability under managed care

The development of managed care in the US has led to fundamental change in the nature of health insurance contracts and in the organization of the medical care delivery. For patients, managed care means accepting restrictions on choice of providers and covered services, in return for lower premiums, lower co-payment, or broader coverage than under traditional fee for service. For providers, managed care establishes risk sharing forms of reimbursement in place of fee-for-service or cost-based reimbursement. Direct controls such as treatment protocols, utilization review, drug formularies etc. narrow the scope of covered services, although patients can purchase non-approved services by paying out-of-pocket. The growth of managed care is a major force driving the restructuring of the delivery system, including horizontal and vertical mergers, alliances and integrated delivery systems, in order to better pool risks, control costs through economies of scale and scope and coordination of care, and compete better for multistate employer contracts while protecting bargaining power.

The growth of managed care, together with the associated changes in provider relationships, has led to new grounds for liability claims. Two types of claim in particular raise issues that are fundamental to the efficiency of the liability and health care systems. First, claims for refusal to pay for care may be brought against the managed care organization (MCO), the physician or the utilization review entity. This raises questions about the appropriateness of changes in treatment norms. Second, in cases alleging negligent treatment by an individual physician, a claim may also be brought against the MCO in addition to the physician. This raises questions of whether managed care plans, which are often insurance entities that contract with health care providers, should be held liable for the negligent care of their contracted providers, if they have exercised due care in screening and selection. These issues are discussed in detail in Havighurst (1995, 1997) and Danzon (1997).

An issue which pervades all cases but is particularly relevant to claims for withholding care is the definition of the standard of care for patients enrolled in managed care health plans. As discussed earlier, due care is traditionally defined as customary care. However, when most patients have comprehensive, fee-for-service insurance, customary care is likely to exceed the social optimum for quantity and some dimensions of quality of care, because of the moral hazard created by traditional indemnity insurance. Managed care can be viewed as a competitive response of insurance and medical markets to the growing demand of consumers and employer/payers for forms of insurance that provide better value for money than traditional indemnity insurance. But if the purpose of managed care is in part to eliminate the excesses and distortions of indemnity insurance, then if courts adjudicate managed care cases using the fee-for-service norms of care, the ability of managed care to reduce the waste of traditional norms will be undermined.

9.1. Claims for denial of coverage

In Wickline v. State, the plaintiff sued MediCal for negligent denial of coverage of an 8-day hospital stay, claiming that the 4-day stay permitted caused the subsequent complications that led to the amputation of her leg. In this case the court ruled for the defense, finding the MediCal restriction was not the cause of her discharge, which was the responsibility of her physician who did not appeal the decision. The court also noted that the discharge was consistent with usual standards of medical practice in the community and hence that MediCal was not culpable of breach of duty. However, the court left open the possibility that third-party payers could be liable for "defects in the design or implementation of cost containment mechanisms", holding it "essential that cost limitation programs not be permitted to corrupt medical judgment". Other courts have used the standard of "medically necessary" care.

A very different conclusion was reached in Fox v. Healthnet, in which a California jury awarded approximately $89m., including $77m. in punitive damages, finding that Healthnet acted in bad faith, breached its contract for care and intentionally inflicted emotional distress through reckless denial of coverage of a bone marrow transplant for treatment of Mrs. Fox's breast cancer. Although the award was drastically reduced in settlement prior to appeal, the potential magnitude and publicity of such awards could

clearly influence HMOs' coverage decisions. Unlike Wickline, Fox did not allege negligence but relied on contract theories commonly used to challenge insurance coverage decisions.

These coverage denial cases raise two issues. First and most problematic, What is the basis for liability for denial of coverage, if any? The courts appear to implicitly assume the existence of an objective and appropriate standard of care defined by medical judgment, which in turn defines appropriate cost containment mechanisms. However, at best medical science can tell us the probability distribution of health outcomes and risks from particular medical treatment. Deciding whether the treatment is worth performing requires comparing the value of the expected outcome to the costs. Valuation ultimately depends on consumer preferences and willingness-to-pay. For private programs, this can be evaluated using willingness-to-pay. For public programs, willingness-to-pay can be defined to include the altruistic willingness-to-pay for others too poor to pay for themselves. Based on this analysis, Danzon (1997) concludes that claims for denial of coverage should be viewed not as negligence claims but as contract disputes, in which the question is: Would enrollees (or similar consumer groups) be willing-to-pay for insurance coverage of this service ex ante, given the cost and expected outcomes? Inevitably, an ex ante willingness-to-pay standard will appear to conflict with the interests of the individual patient once sick, who would then want coverage of all services that offer any positive expected benefit. However, to achieve an efficient standard of liability for coverage disputes, courts must ignore the ex post or patient-specific private optimum and focus on the ex ante or group optimum, which also approximates the social optimum (ignoring tax distortions). But it is the ex post or private patient optimum that underlies the traditional norms of indemnity insurance.

This contractual approach to coverage disputes would permit the standard of care to vary, depending on the type of plan, the premium and explicit and implicit contractual terms. If instead, courts apply a uniform standard of "medically necessary" care to all plans, health plans will be constrained in their incentives to compete by developing innovative, more cost-effective patterns of care and to differentiate their product offerings to cater to heterogeneous consumer preferences.

The second issue is, In the event of failure to pay for services in conformity with the contract, who should be liable – the health plan, the doctor, the UR agency or all three, on grounds of joint and several liability? Transactions costs considerations indicate that liability should be placed only on the health plan, not the individual provider or UR agency, since it is the plan that defines the contract, operates or contracts for the UR controls, and ultimately bears the financial risk of paying for the contracted services within the premiums paid. If liability is placed on individual physicians or UR agencies, they are likely to seek contracts of indemnification from the health plan, hence it is more efficient to place liability directly on the plan.[16]

[16] Of course in a provider-sponsored MCO, liability in coverage disputes would be on the provider group in their role as plan sponsor, not as a provider of care.

9.2. MCO liability for negligent care

Since managed care plans typically restrict enrollees to the network of selected providers, plans are required to use due care in selecting and monitoring participating providers and may be held liable for negligence in performing these credentialing functions. Similar liability for negligent credentialing already applies to hospitals with respect to their credentialing of staff physicians, including independent contractor physicians with admitting privileges.

A separate and far more contentious issue is whether an MCO should be liable, under theories of vicarious liability or ostensible agency, for the negligence of its contractor physicians, assuming that the MCO has exercised due care in credentialing. Proponents argue that patients look to HMOs as providers of care, in part as a result of the HMO's own promotional material. Havighurst (1995, 1997) argues that the default position should be enterprise liability of health plans (for POS plans, liability would only extend to torts of affiliated providers). He would offer MCOs the option of contracting for other allocations of responsibility. His argument is that "MCOs, although in control of many levers that can affect the quality of care for better or for worse, are not, in the eyes of the law, routinely answerable for poor quality. Enterprise liability is the logical legal culmination of the shift to de facto corporate responsibility that is revolutionizing American medical care" [Havighurst (1997, p. 588)].[17]

The argument against holding MCOs liable for the negligence of their contracted physicians is that, for loose networks such as IPA and POS plans, the plan lacks the information and authority to control the details of care delivery, which remains in the hands of individual providers. Danzon (1997) argues that liability for negligent performance should therefore remain with these individual providers, if the MCO has adequately met its obligation to screen participating providers. To add the MCO as a defendant in claims alleging negligence by the provider, simply adds another deep pocket defendant, which may distort the outcome of claims, without adding useful deterrence. Exposing MCOs to liability for the negligence of their contracted providers is likely to lead them to select more restrictive networks and exercise tighter control, which appears contrary to the preferences of consumers who increasingly choose plans with broad networks and POS options.

Ultimately, the issue is whether it is practical and sensible to distinguish between coverage decisions and negligent performance, conditional on the coverage decision. Danzon assumes that this distinction can be made and is important, hence argues for plan liability for coverage decisions but provider liability for implementation of treatment. By contrast, Havighurst subsumes these different dimensions into a broad notion of "quality", notes that plans do intervene in some dimensions of this broadly defined quality, and hence concludes that they should be liable for all aspects, including treatment.

[17] He also argues that placing liability on health plans would lead them to take more effort to define the standard of care in their contracts with patients.

9.3. The ERISA pre-emption

The federal Employer Retirement Income Security Act (ERISA) preempts state laws to the extent that they "relate to" an employee benefit plan that is subject to ERISA, which includes all self-insured employer plans. This has been interpreted to bar tort claims for denial of coverage by employees against HMOs where the HMO coverage is sponsored by a private employer.[18] Employees may receive compensation of some money damages if, for example, they could show an administrator's misconduct, but compensation for pain and suffering and punitive damages would not be authorized. The extent of the ERISA preemption has been uncertain, with the trend toward gradual erosion through court decisions and explicit statutory changes.

The ERISA preemption is consistent with the contractual view of employee benefit plans. Economic analysis concludes that, in the long run, the costs of employer-provided health insurance are borne by employees through lower wages and that employers have incentives to design such plans to maximize the utility of covered employees, since this minimizes the money wage or other benefits that must be offered to attract a given workforce. Given this interpretation, it makes no sense to permit individual employees to sue for denial of coverage, except where contractual commitments have been breached. Efficiency and equity argue for eliminating the inconsistencies between the liability exposure of ERISA-protected plans and non-ERISA plans, the question is which rule should prevail. Some current legislative proposals would effectively eliminate the ERISA preemption by granting patients statutory rights to sue HMOs and possibly their employers. The analysis outlined above suggests that, in the absence of a well-defined, contractual approach to defining responsibilities for paying for care, such an extension of liability could seriously undermine the ability of HMOs to eliminate the waste that was embodied in customary care.

10. Non-US experience

10.1. The UK

Medical negligence and its costs emerged as a public policy issue in the UK in the 1980s, following several years of rapidly rising claim costs and premiums. In 1996, medical negligence is estimated to have cost the NHS in England £235m., with an estimated rate of increase of 17–25 percent per annum, and is again a major concern [Dobson (1998)].

The liability of medical providers in the UK derives from the same common law origins as in the US, but with significant difference in detail. First, in the UK cases are decided by common law judges rather than juries. Second, tort awards are reduced by

[18] The employer sponsor in Fox was a public school district and hence was not protected by ERISA.

the amount of compensation available from social insurance programs and NHS medical benefits, and awards for pain and suffering are more modest. This full collateral source offset, which also occurs in Canada, New Zealand, Sweden and most European countries, is a major factor contributing to the apparent differences in costs of medical negligence in different countries. Third, lawyers traditionally are paid by the hour and are not permitted to take contingent fees. The English rule allocates the legal costs of both sides to the losing party. In theory, this loser-pays-all rule could significantly deter risk averse plaintiffs from bringing claims. In practice, most medical negligence plaintiffs in the UK receive Legal Aid and the English rule has not been applied to Legal Aid; however, the hourly rates paid by Legal Aid may not fully compensate some lawyers for the opportunity cost of their time. Whether on balance the incentives of patients and lawyers to bring claims are too high or too low, relative to the social optimum, remains an empirical question.

The data on injuries and claims in the UK are very limited. There has been no comprehensive study of the incidence of iatrogenic injuries and negligent injuries. Although some analysts have extrapolated the findings of the Harvard study for New York state to estimate the number of negligent injuries in the UK [Smith (1990)], such estimates are tentative at best, because they assume the same rate of iatrogenic injury and negligence in the UK as in New York. As noted earlier, in the New York study the count of iatrogenic injuries depended on the standard of care implicitly assumed by the reviewers, case mix and severity of hospital admissions, and the count of negligent injuries depended on the reviewers beliefs about legal standards. It is a big leap to assume that rates of adverse events and negligent injuries defined in this context-specific manner would be the same in other countries that differ from New York in their norms of care, case and severity mix of hospital admissions, and legal standards.

There are no comprehensive data on the level and trends of negligence claims for the UK. Based on a careful study that pieced together information from several sources [Ham, Dingwall et al. (1988)], it appears that from the mid-1970s to the mid-1980s the rate of increase in number of claims and size of awards was at least as rapid in the UK and Canada as in the US. Nevertheless, because these countries started from a lower base, by 1987 doctors in the US were still five or six times more likely to be sued than doctors in Canada or the UK, and awards for comparable injuries were several times larger in the US [Danzon (1990a, 1990b)]. This overstates the difference in real compensation to patients, because the attorney's fee (usually one third) is subtracted from the award in the US and because medical costs are borne by public health care systems in the UK and Canada.

Although medical negligence costs to the NHS are reportedly growing at 17–25 percent per annum, the basis for these estimates is unclear. Also unclear is the relative contribution of increased frequency of claims, increased severity of claims or simply increased rates of treatment; whether the reported figures refer to costs paid out or costs accrued in a given year; and how far the apparent cost growth reflects a shift from pay-as-you-go accounting for hospital trusts to accrual accounting [Towse and Danzon (1998)].

Recent changes in the locus of liability and insurance responsibility in the NHS have complicated the tracking of claim trends. Prior to 1988, all NHS doctors were required to join one of the two medical defense organizations (MDOs). In the mid-1980s, as claim costs and premiums rose, commercial insurers entered the market offering lower rates to GPs and other low risk specialties, undermining the traditional community-rated premiums of the MDOs.[19] The ensuing price war for low risk specialties threatened significant premium increases for high risk specialties, including orthopaedic surgeons, obstetricians, etc., particularly in the higher risk urban areas, which in turn threatened the NHS policy of equal net incomes. The outcome was that the NHS instructed Health Authorities to assume full responsibility for all new and existing claims against employed staff, up to a cap of £300,000 (including legal costs), with responsibility for claims from pre-1990 practice to be met partly by the MDOs. GPs continued to have their subscriptions (premiums) fully refunded ex post through their expenses. With the formation of trust hospitals, the responsibility for negligence claims devolved from the health authorities (HAs) to the trusts.

In effect, the UK thus established a form of enterprise liability, in which either the HAs or the trusts are responsible for all liability arising out of the practice of their consultants on NHS business (consultants must continue to provide for their own cover for their private business). This experience is thus of considerable interest to the enterprise liability debate in other countries. Since claimants had usually sued the HA as well as the doctor, on grounds of vicarious liability, the practical effect was to reduce the doctors' role in claims settlement and their concern over premium rates. The profession opposed NHS enterprise liability on grounds that it would lead to settlements that ignored effects on professional reputation and lead to greater NHS management of clinical activity and interference with medical judgement – arguments often made by physicians in the US who oppose enterprise liability. In practice, the HAs are probably too far removed from the delivery of care to exercise well-informed risk management and loss control. Thus the net effect of shifting liability from the individual consultants to the HA would probably be to reduce deterrence.

Shifting enterprise liability from HAs to individual trust hospitals offers greater opportunity for realizing the potential risk-management benefits of enterprise liability. However, trusts' incentives to engage in risk management and loss control depend on the extent to which they bear the costs of negligent injuries. In 1995, the NHS litigation authority was set up to manage the Clinical Negligence Scheme (CNS) for trusts. This is a voluntary reinsurance arrangement, in which participating trusts pool their loss

[19] The MDOs are technically not insurance companies, hence are not subject to the same reporting, reserving and solvency requirements as insurance companies and are under no legal obligation to pay claims against their member doctors. In practice, MDOs perform the same functions as insurance companies. However, their looser reporting and reserving requirements may enable the MDOs to operate and set prices on a pay-as-you-go basis rather than an accrued cost basis. The special legal status of MDOs may also reduce the contestability of the liability insurance market in the UK, Canada and other countries in which MDOs operate. Such entry barriers reduce competitive pressures for experience rating and permit community rating to survive.

above an excess threshold ranging from £10,000 to £500,000, depending on the trust's size and activities. All but 6 of the first 226 joiners chose the lowest excess, typically £25,000. Since there is no participation in losses above the retention and premiums are apparently not experience rated, this reinsurance arrangement significantly undermines trusts' incentives for loss control, although the remaining retained risk should leave some incentive (the mean settlement is around £50,000). However, a further critical question is the actual incidence of liability costs. If trusts face a very inelastic demand for their services, because of local market power and/or long term contracts with HAs, then those trusts that incur relatively high liability costs can simply pass on these costs as higher charges to payers. Even if the CNS reinsurance premiums were experience rated, the deterrence incentives that flow from enterprise liability in the NHS are likely to be weak if competition is weak in the market for hospital services, such that trusts can simply pass through their liability costs to payers, as higher charges and/or reduced services provided for fixed budgets. Thus in the case of the NHS, replacing the liability of individual consultants by enterprise liability of HAs or trusts has probably undermined deterrence, by reducing both the monetary and non-monetary incentives of doctors to take care, and because weak competition in the market for hospital services permits trusts to pass on rather than internalize their liability costs.

The UK experience with claims for denial of payment for care is also of great interest as the US wrestles with the same issues in the context of managed care. The UK courts have consistently supported the HAs in cases where treatment has been denied on grounds of shortage of funds, provided that the decision making process is reasonable [for example, Ham (1998)]. Thus the UK courts defer to payers to make substantive resource allocation decisions of their limited budgets and to define the process, and so far have protected them from liability. It remains to be seen whether US courts will similarly recognize that HMOs and other insurers must operate within budget constraints and respect their autonomy in both the process and the substantive outcomes of the inevitable resource allocation decisions. A key difference appears to be that UK courts believe that money saved in denying care to one patient will be spent on another patient, whereas US courts may be led to believe that money saved in denying coverage goes into the pockets of the HMO shareholders. Both naïve views are probably too extreme: UK HAs are not necessarily perfect agents for consumers, whereas US HMOs are under competitive pressure to deliver value for money to consumers. Nevertheless, the naïve views may prevail over these economic views, permitting UK courts to adopt more appropriate, ex ante social criterion in these coverage cases, rather than the ex post, individual patient criterion that US courts may be pressured to adopt.

10.2. The Swedish patient compensation insurance (PCI)

The PCI was established in 1975 by voluntary contract between medical providers and a consortium of insurers, to preempt the threatened statutory expansion of tort liability

in order to make compensation available to a larger number of patients.[20] Although patients retain the choice to sue in tort under traditional negligence rules, tort claims have been extremely rare since the PCI was established. Key features of the Swedish model include the elimination of the concept of negligence or fault and decoupling of patient compensation from deterrence or any sanctioning of providers. Patient compensation is provided by the Patient Compensation Insurance (PCI), while the discipline of medical providers is handled by the Medical Responsibility Board (MRB). No information is transmitted between them, in order to obtain the doctors' cooperation with the PCI. The PCI is administered by a consortium of insurers, with appeal to a special advisory panel and ultimately to arbitration.

The Swedish model has attracted interest in other countries because of its relatively low budget cost, low overhead rate and widespread acceptance by medical providers. In the early 1990s, claim frequency stabilized at about 21 per 100 physicians per year, compared to 13–16 claims per 100 physicians in the US. The higher rate in Sweden is not surprising, given the lower costs of filing, somewhat broader criteria of compensability and less reason to oppose payment of marginal claims (see below). Roughly 40 percent of these claims receive compensation in both countries. But the PCI costs roughly $2.38 per capita, or 0.16 percent of health care costs in Sweden, whereas medical malpractice insurance premiums are about 1 percent of higher health expenditures in the US. Thus the per capita budget cost of the Swedish PCI appears to be roughly one tenth of US. malpractice premiums. Administrative overhead is 14–18 percent of total PCI premiums, compared to roughly 60 percent in the US. This low overhead rate is often cited as evidence of the potential savings from eliminating negligence in favor of a no-fault (causation-only) rule of compensability for medical injuries [Weiler (1991)].

However these inferences are based on a misunderstanding of the PCI. The low budget cost of the PCI, despite the higher claim frequency, reflects primarily two factors. First, the collateral offset rule shifts most of the wage loss and medical expense of iatrogenic injuries to other social insurance programs. In fact, such cost-shifting does not reduce and may actually increase the social costs of injuries, by undermining cost internalization which undermines general and specific deterrence. Because the economic loss related to iatrogenic injury is largely shifted to other social programs, the PCI budget vastly understates the true cost of compensating iatrogenic injuries in Sweden.

Second, awards for noneconomic loss are below those in most other European countries and roughly one tenth of those in the US.[21] The PCI can pay low awards without driving patients to select the tort option because Sweden's tort system offers even lower

[20] Only roughly ten patients a year received compensation through the traditional negligence-based tort liability system. One alleged obstacle was the reluctance of physicians to testify on behalf of plaintiffs, as was required by the custom-based standard of care. Risk-averse plaintiffs may also have been unwilling to pay the legal fees in the absence of contingency arrangements or legal aid.

[21] In 1987, the mean payment for noneconomic loss under the PCI was $3,800, the maximum was $117,070. Nevertheless payments for noneconomic loss account for roughly 74 percent of total PCI payments, because economic loss is heavily covered through collateral sources.

payments than the PCI, with lower probability of success.[22] Thus the PCI offers plaintiffs an expected payoff that at least matches their expected tort recovery, net of costs, in order to deflect tort claims. However, other countries that have more generous tort systems could not adopt a voluntary contractual alternative such as the Swedish model and expect to realize costs as low as in Sweden, unless they also adopted significant tort reform. Indeed, if a Swedish-style PCI were offered as a voluntary alternative to tort in the US, it would not offer significant cost-savings relative to those that can be realized through out-of-court settlement, which already provides a voluntary, contractual alternative that operates in the shadow of the tort system.

The PCI's low overhead percentage is not the result of using a causation-only test for compensability. Although the PCI is often called no-fault, this is misleading. From the patient's perspective, the criteria of compensability are quite similar to a traditional negligence rule based on customary practice. Under the PCI, an injury is compensable if (1) it occurred with "substantial probability" as a direct consequence of medical intervention, and (2) either the treatment was not medically justified or the injury could have been avoided by performing the treatment differently. Thus although the terminology of fault and negligence have been eliminated, compensation requires some notion of "error". Adverse outcomes caused by medical care are explicitly excluded, if the treatment was medically justified.

But from the provider's perspective, the PCI is both no-fault and no-liability. The PCI eliminates all reference or inquiry into fault or negligence, does not require the patient to identify a particular provider who failed in a duty of care, and entails neither financial nor reputational consequences for individual providers. This no-liability scheme bears no resemblance to the proposed strict enterprise liability of hospitals [Weiler (1991)], since the strict enterprise liability proposal would place the liability for paying damages on the defendant hospital, in order to promote deterrence.

The low PCI litigation percentage reflects several features that are unlikely to be acceptable in the US. These include the elimination of all link between patient compensation and provider liability and deterrence, and the modest level of patient rights, compared to a US tort plaintiff (although not necessarily compared to a Swedish tort plaintiff). Under the PCI, because physicians are not liable and have no personal stake in the outcome, they generally cooperate rather than opposing compensation of an injured patient. Patients have little to gain from appeal to the review panel or to arbitration. Both are closed to the press and public, and the panel has ruled in favor of the insurers in 90 percent of cases.[23]

Thus the low litigation rates reflect primarily the fact that neither plaintiffs nor physicians have strong incentives to oppose or appeal the insurers' adjudication of claims. Other contributing factors are the simple claim filing process; administration

[22] Among other obstacles, plaintiffs allegedly have difficulty obtaining the expert testimony required to support a claim for negligence under the custom-based negligence rule.

[23] Since 1992, major decisions of the review panel and all arbitration decisions are published.

by a monopoly consortium of insurers, which eliminates insurers' incentives to vigorously oppose plaintiff claims or to experience rate premiums; and lack of competition and provider-specific accountability for costs in the health care system, which makes providers more willing to tolerate flat-rated premiums, despite significant geographic differences in claims experience. [24]

The deterrence function, to the extent that it exists, was assigned to the Medical Responsibility Board. Patients can file a claim with the Medical Responsibility Board if they feel that their treatment was negligent or contrary to the statutory code of medical practice. They bear their own filing costs and receive no compensation. Providers may receive a reprimand or warning, but this has no financial consequence and probably at most a minor reputation effect. There are roughly 6 MRB claims per 100 physicians per year, of which one in 6 receives some sanction. Thus the ratio of MRB sanctions to paid PCI claims is less than 1 in 10 – a rough measure of the loss in potential deterrence that results from decoupling compensation from medical discipline [Danzon (1994a)]. Although the PCI database on iatrogenic injuries might in theory be used to identify and control persistent sources of risk, in practice the information collected is insufficient. Moreover, although clinics and hospitals are informed about their claims experience, the responsible individuals and sometimes even the nature of the injury are not identified.

Thus the main lesson from the Swedish PCI experience is that a sufficient and possibly a necessary condition for low overhead costs is to forego all links between compensation and deterrence. Whether this results in more iatrogenic injuries with social costs that outweigh the reduction in litigation costs is an unanswered empirical question. The answer would probably differ across countries, depending on the costs and benefits of their tort systems and on the costs and effectiveness of other mechanisms for quality control.

10.3. The New Zealand accident rehabilitation compensation and insurance scheme (ACS)

The New Zealand ACS was established in 1972 as a comprehensive no-fault compensation system for victims of "personal injury by accident", including "medical misadventures". The establishment of the ACS followed the elimination of traditional tort rights for such injuries, in contrast to the Swedish PCI, where tort actions remain an option. Claims are administered by the ACS, with appeal to a special ACS Authority. Compensation was set at a relatively high percentage of wage loss for workers, plus scheduled lump sum payments for noneconomic loss. Medical costs were borne by the

[24] The PCI is financed by premiums paid by the county councils, who are responsible for financing and provision of the public health care system in Sweden, and by private physicians, dentists and other paraprofessionals. For each provider category, premiums are assessed on a flat per capita basis, regardless of claims experience.

National Health Service (NHS), except that the ACS paid directly for services in private hospitals, co-payments and services not covered by the public system.

Between 1975 and 1989 total expenditures for all types of injury covered by the ACS grew at a nominal rate of over 20 percent a year, or roughly 6 percent a year after adjusting for inflation, which exceeds the average rate of increase of US malpractice premiums over the same period. For 1985–1997, expenditures rose at an annual real rate of 8 percent [NZBR (1998)]. However, these trends are not strictly comparable to trends in malpractice costs in other countries because the ACS data include all injuries. The ACS did not track iatrogenic injuries as a separate category until 1993. Financing is through payroll and general taxation, with no separate assessment of medical providers.

Concern over the rapid increase in ACS costs, the inequity of the incidence of costs (low risk employers and drivers subsidizing high risks) and the neglect of injury prevention led to significant changes in the 1992 Accident Rehabilitation and Compensation Insurance Act. This 1992 Act redefined medical misadventure. The problems under the original ACS and the changes adopted are instructive.

"Personal injury by accident" was broadly defined by the original ACS statutes to include "physical and mental damage caused by medical, surgical, dental and first aid misadventure". The intent was to exclude illness and normal risks of medical care but to include medical injuries that fall outside the realm of normal risk, including those caused by negligence or with either very low probability or unexpected severity. Difficulties in implementing this definition led to various proposals for change over the years, including use of ICD-9 definitions of injuries, or extending the system to include all incapacity. The broad range of reform proposals may reflect the conceptual problem faced by any compensation scheme for iatrogenic injuries that is not focused on deterrence, of providing an equitable justification for compensating some victims but not others in similar condition but from other causes.

The 1992 ACS reform provides a statutory definition of medical misadventure as "personal injury resulting from medical error or medical mishap". "Medical error" is "the failure of a registered health professional to observe a standard of care and skill reasonably to be expected in the circumstances". "Medical mishap" is an adverse consequence of treatment that is both rare and severe. An injury is "rare" if it has less than a 1% probability of occurring. An injury that would normally be rare may not be rare given the particular circumstances of the patient, if this greater risk was known to the patient. An injury is severe if it results in death, hospitalization for more than 14 days, significant disability lasting more than 28 days or qualifies for ACS's Independence Allowance. Medical mishap specifically excludes delayed abnormal reactions and complications of procedures. Injuries related to lack of informed consent, misdiagnosis or treatment omissions are compensable only if they result from negligence.

The 1992 reforms require the ACS to pay for all medical costs incurred by compensated victims, effectively restoring the traditional collateral source rule, in the interests of accountability and to increase the ACS's incentives for loss control.

The New Zealand ACS has often been acclaimed for its low overhead costs (less than 10 percent of total expenditures) and prompt payment of compensation. Traditionally,

this partly reflected the ACS's practice of accepting the majority of claims as filed, relying largely on physicians as gatekeepers to certify that a claim is a "personal injury by accident", and the simple claims adjudication procedures. Data collection costs were kept low, with little detailed information on the causes of injuries. Thus like the PCI, the ACS has maintained low overhead by simply adjudication and data systems and by severing all links between compensation and deterrence, which eliminates the incentives of those causing injuries to oppose the compensation. Whether such a strategy of skimping on overhead is "penny wise but pound foolish", leading to higher real social costs of injuries, is an important but unanswered empirical question.

Under the 1992 reforms a separate Medical Misadventure Unit of the ACS was established to handle medical injury claims. It has ultimate authority on all claims, but seeks advice on complex claims (including all medical error claims) from a three-person advisory committee (a health professional from the relevant specialty, a lawyer and a lay person) drawn from a pool of non-ACS people. Both parties are given 15 working days to comment on the preliminary findings of the advisory committee, which may meet up to 3 times. Roughly 45 percent of claims have been accepted; of these, roughly 86 percent are based on mishap (no fault) and 14 percent are based on error (negligence) [ACS (1996)]. Claimants may appeal for review of the decision by an independent office within ACS. Providers may appeal only those medical error claims associated with negligent failure to obtain informed consent or misdiagnosis.

Although the 1992 legislation provides for the Medical Misadventure Account to be funded by premiums levied on health care professionals, so far this has not been implemented and the account remains funded out of general and payroll taxes. Imposing experience-rated premiums on individual providers would expose them to great risk that is largely beyond their control (since only 14 percent of claims are due to negligence) which would be inequitable and inefficient, particularly since providers are paid fixed fees that are not risk-adjusted for patient mix. Those compensated claims that are deemed attributable to negligence are reported, but the ACS has no record of whether disciplinary proceedings were taken.

The original New Zealand ACS structure illustrates pitfalls to be avoided rather than a useful prototype that other countries might adopt. The original definition of a compensable event was difficult to implement. The 1992 reforms clarified the definition of medical mishap, adding explicit reference to medical error and negligence, with associated possibilities for disciplinary actions. The low reported administrative costs of the ACS reflect relatively low investments in determining the causes of injuries, limited data collection and little attempt at deterrence, which may have contributed to the rapid escalation of total costs. These comments apply to injuries in general, because medical injuries were not identified prior to 1992. If the proposed introduction of experience-rated premiums for medical providers were implemented, providers would have more incentive to oppose claims and litigation costs would likely increase. A more accurate measure of the true overhead cost of an accident compensation scheme would include not only the reported overhead but also the unmeasured deadweight loss from unnecessary injuries and inappropriately compensated claims. This is not observable, but is

likely to be higher for both the PCI and the ACS than for tort-based systems which incur higher reported overhead costs with the purpose of deterring inappropriate injuries.

11. Concluding comments

The basic rationale for medical malpractice liability is to improve provider incentives for safety, assuming that asymmetric information leads to market failure in medical markets. The evidence of a significant rate of negligent injury, invalid claims and physicians' preference for insurance policies with minimal explicit co-payment or experience rating indicate that the efficiency of the malpractice system is severely constrained by imperfect information on the part of courts, doctors, patients and liability insurers with respect to appropriate care and legal standards. Just as imperfect information undermines the efficient functioning of the market, imperfect information undermines the efficient functioning of the liability system. The fundamental problem is that changing the liability rule does not correct the information asymmetry.

Nevertheless, the evidence shows that the worst criticisms of the malpractice system as a random lottery are unfounded. Negligent injuries are more likely to lead to claims and receive payment than non-negligent injuries, many invalid claims receive zero or low payments, and payments are strongly related to the loss incurred by the patient. Malpractice also does not appear to be a major driver of health care costs. Malpractice premiums are only roughly one percent of total health care spending. Defensive medicine has not been accurately measured or distinguished from insurance-induced overuse of care. To the extent that managed care internalizes to providers the costs of unnecessary care, defensive medicine is likely to become a less pressing issue.

The fault-based system of liability for medical injuries may be worth retaining if the benefits, in terms of injuries deterred, exceed the excess costs of litigating over fault and defensive medicine. A full cost-benefit evaluation is not possible given the available data, but rough calculations suggest that a positive net benefit is plausible. Nevertheless, overhead costs are high. Certain incremental reforms could make the system more cost-effective for purposes of both deterrence and compensation, in particular, scheduled awards for non-economic loss; redefinition of the standard of care to reflect managed care and heterogeneity of health plans; and possibly some forms of ADR. Freedom to contract out of the system, for alternative rules of liability (no provider liability, liability only for gross negligence, etc.) for alternative benefit levels or for arbitration of disputes, could be encouraged as part of the health insurance contract. Freedom for providers to contract for various forms of enterprise liability could offer savings if adopted voluntarily. For the US, the key issue going forward is to find the best basis for holding managed care plans to their contractual commitments to care, without undermining their ability to continue to search for better ways to define and implement appropriate care.

There is no consensus on whether switching from a fault-based rule (possibly more clearly defined) to a no-fault basis for compensation would offer significant savings in litigation costs. The evidence from Sweden and New Zealand suggest that other factors

contribute to their low overhead costs. If the deterrence benefits are considered too uncertain to warrant retaining malpractice liability of individual medical providers, then there is no strong case for a special, tax-financed compensation program for victims of iatrogenic injury. The adequacy of compensation for iatrogenic injury is simply part of the broader question of the efficiency and equity of the existing network of private and social insurance programs.

Appendix

A1. Optimal liability rules

This model is formulated in terms of medical injuries but could apply in any market context where consumers may be injured by product failure which depends only on the care taken by producers. State-dependent utility of patients and physicians is assumed, because personal injury plausibly affects the patient's utility of income and liability claims impose uninsurable time and reputation costs on physicians.

Assume that expenditure on prevention affects the probability but not size of loss, and each patient buys just one unit of medical care. Patients can buy first party and/or physicians can buy liability insurance, with perfect experience rating and a proportionate loading charge. The following notation is used:

$V(B)$ = patient's utility of initial wealth, $V' > 0$, $V'' < 0$,
$p(r)$ = probability of injury, $p' < 0$, $p'' > 0$,
r = quality (prevention per unit of service),
s = price of services,
$c(r)$ = production cost per unit, $c' > 0$, $c'' < 0$,
L = monetary loss to patient if injury occurs,
M = first party insurance coverage bought by patient,
γ_p = premium rate per dollar of first party coverage, where $\gamma \geqslant 1$ is the loading charge,
$h(p)$ = patient's perception of p, $h' > 0$, $h'' < 0$,
$U(A)$ = physician's utility of initial wealth, $U' > 0$, $U'' < 0$,
D = damages paid by the physician if a loss occurs,
Q = liability insurance coverage bought by physician, $Q \leqslant D$,
λ_p = premium rate per dollar of liability coverage, where $\lambda \geqslant 1$ is the loading charge,
μ = Lagrange multiplier.

Subscript '0' denotes the state in which an injury occurs. Subscript '1' denotes the state in which no injury occurs. Subscripts 'f', 's', and 'n' denote first party, strict and negligence liability, respectively. Initially, patients are assumed to be fully informed.

A1.1. First party liability (caveat emptor)

If patients are fully informed and markets are competitive, the physician chooses the level of safety (r) and product price (s) to maximize expected utility of patients, $E(V)$, subject to maintaining an opportunity level of utility, U^c, determined by the physician's alternative use of time.[25] Patients select first party insurance coverage (M), given the supply price per dollar of coverage γ_p. Informed markets thus solve the following optimization problem:

$$\max_{M,s,r} \phi = (1-p)V_1\big[B - s - \gamma_p M\big] + pV_0\big[B - s - \gamma_p M - L + M\big] + \mu\big\{U[A + s - c(r)] - U^c\big\}. \tag{A.1}$$

Maximization with respect to M, s and r yields

$$\gamma \overline{V}' = V_0', \tag{A.2}$$

$$\overline{V}' = \mu \overline{U}', \tag{A.3}$$

$$c' = -p'\left[\frac{V_1 - V_0}{\overline{V}'} + \gamma M\right], \tag{A.4}$$

where $\overline{V}' = (1-p)V_1' + pV_0'$.

Equation (A.2) may be written

$$\frac{V_1'}{V_0'} = \frac{1-\gamma_p}{\gamma - \gamma_p} \leqslant 1 \quad \text{as } \gamma \geqslant 1.$$

This is the familiar result, that optimal coverage does not fully equate the marginal utility of income if the insurance premium contains a proportionate loading charge ($\gamma > 1$). Equation (A.4) shows that if injury entails irreplaceable loss ($V_1 > V_0$), optimal prevention (r_f^*) may exceed the optimal level with risk neutrality ($c' = -p'L$), even with full insurance of any monetary loss.

A1.2. Strict third party liability

Under a rule of strict third party liability the physician pays for all iatrogenic injury. An omniscient benevolent dictator would choose the damage award (D), physician's liability insurance coverage (Q), prevention (r) and product price (s) to maximize

[25] If the physician has monopoly power, U^c includes some rent but the structure of the problem is not affected. The model ignores other arguments in a physician's utility function, such as prestige and ethics.

the patient's expected utility, subject to maintaining the opportunity level of utility for physicians U^c:

$$\max_{D,Q,s,r} \phi = (1-p)V_1[B-s] + pV_0[B-s-L+D] + \mu\{(1-p)U_1[A+s-c(r)-\lambda_p Q] + pU_0[A+s-c(r)-\lambda_p Q-(D-Q)] - U^c\}. \tag{A.5}$$

Maximization with respect to D, Q, s and r yields

$$V_0' = U_0', \tag{A.6}$$

$$\lambda \overline{U}' = U_0', \tag{A.7}$$

$$\overline{V}' = \mu \overline{U}', \tag{A.8}$$

$$c' = -p\left[\frac{V_1 - V_0}{\overline{V}'} + \frac{U_1 - U_0}{\overline{U}'} + \lambda Q\right], \tag{A.9}$$

where $\overline{U}' = (1-p)U_1' + pU_0'$.

Equations (A.6) and (A.8) together imply:

$$\frac{V_1'}{V_0'} = \frac{U_1'}{U_0'} = \frac{1-\lambda_p}{\lambda - \lambda_p} \leqslant 1 \quad \text{as } \lambda \geqslant 1.$$

The optimal tort damage award, D^*, provides the level of insurance patients would choose to buy voluntarily, but with the loading factor of the physician's liability insurance. Thus $D^* \gtrless M$, as $\lambda \gtrless \gamma$. The optimal safety level (r_s^*) implicitly defined by Equation (A.9) may exceed or fall short of (r_f^*) under first party liability as $(U_1 - U_0)/\overline{U}' + \lambda Q \gtrless \gamma M$.

If $\lambda = \gamma = 1$ and the physician incurs no uninsurable liability loss, then $r_s = r_f^*$. In the more realistic case of $\lambda > \gamma$ and $U_1 > U_0$ because of time and reputation costs of suit, the optimal level of prevention is higher under strict liability.

A1.3. Negligence

Under a negligence rule, the physician is liable only if he fails to meet the due care standard, r_n^*, and the patient is injured. The social welfare function is given by Equation (A.1) for $r < r_n^*$, and by Equation (A.5) for $r > r_n^*$. If either U is state-dependent or $\lambda > 1$, the physician is not fully insured against the loss. The social welfare function is discontinuous at r_n^* because of the costs of suit to physicians.

A first best solution can be achieved by setting D at D^*, the optimal compensatory award under strict liability, and $r_n^* = r_f^*$. The physician's decision problem is then to choose Q, s and r to maximize $E(V)$, subject to $E(U) > U^c$ and subject to the penalty $D = D^*$ if $r < r_n^*$. But this private objective function is identical to the social welfare

function, i.e., it is a discontinuous function equal to Equation (A.1) for $r > r_n^*$ and equal to Equation (A.5) for $r < r_n^*$. If either $\lambda > \gamma$ or $U_0 < U_1$, the physician's incentive is to choose $r = r_n^*$, i.e., to be non-negligent and hence have no demand for liability insurance. On the other hand, if $1 = \lambda < \gamma$ and U is not state-dependent, then physicians would choose $r < r_n^*$, i.e., would choose to be negligent and to insure the resulting losses. As a practical matter, this case can be ignored.[26]

A2. Imperfect information

The analysis so far has shown that if consumers are fully aware of risks, first party liability and negligence are equally efficient and superior to strict liability, when shifting liability imposes uninsurable losses on defendants.[27] But if consumers underestimate risks, under first party liability they buy too little insurance and non-optimal safety. Spence (1977) shows that under strict liability, a first best solution with respect to compensation and prevention can be achieved by means of a two-part penalty. A compensatory award equal to D^* is paid to victims. A fine, paid to the state initially but refunded as a subsidy to the hazardous activity, is set equal to $(1-h')[(V_1 - V_0)/\overline{V}']$, where $h(p)$ is the consumer's perception of p and $(V_1 - V_0)/\overline{V}'$ is the dollar measure of loss due to injury, or the willingness to pay for injury reduction.

With a risk averse defendant and incomplete insurance, provided the standard of care is correctly set at r_f^*, the fine necessary to achieve compliance is less under negligence than under strict liability because of the discontinuity of the payoff function.[28] The physician will choose to meet the standard provided:

$$\frac{U_1 - U_0}{\overline{U}'} + \lambda Q \geqslant \frac{V_1 - V_0}{\overline{V}'}\left(1 - \frac{h'}{p'}\right) + \gamma M.$$

Thus if the load on liability insurance is at least as great as the load on first party insurance ($\lambda \geqslant \gamma$), a fine over and above the compensatory award paid to victims is not necessary to induce compliance with a negligence standard, if the compensated cost of suit to physicians $(U_1 - U_0)/\overline{U}'$ exceeds the distortion in market incentives due to consumer misperceptions $[(V_1 - V_0)/\overline{V}'](1 - h'/p')$. Since the fine-subsidy mechanism is presumably costly to administer, this is an added attraction of a negligence rule over strict liability.

[26] If $\lambda > \gamma \geqslant 1$, it might be optimal to provide compensation through first party coverage and impose a liability fine on physicians to achieve optimal deterrence. Enforcement would depend on subrogation actions by the patient's first party insurer against the physician or his liability insurer.

[27] This analysis ignores costs of adjudicating claims, which would probably be highest for strict liability, lowest for first party liability.

[28] Note that there is a range within which the due care standard can fall and still induce compliance.

A3. The demand for liability insurance

It has been shown that, under a negligence rule with the standard of care and rule of damages optimally defined, physicians have incentives to be non-negligent. Hence there should be no demand for liability insurance.

This argument presupposes that courts enforce an efficient due care standard with perfect accuracy, and that this is known to physicians and patients. The demand for liability insurance can arise out of either Type 1 or Type 2 errors by the courts, or penalties insufficient to offset consumer misperception of risk. If courts set the standard too high, holding physicians liable for some injuries where the cost of prevention exceeds the expected benefits (Type 2 errors), it is cheaper for the physician to insure than to prevent these injuries, and this is socially optimal. But if victims or courts also commit Type 1 errors, failing to file or to find liability in all true instances of negligence, or if liability payments are too low, then it is cheaper for the physician to insure than to avoid some instances of negligence, and this is not socially optimal.

These two cases are illustrated in Figure 1. The curves labeled FF and SS show the marginal social benefits of care per patient encounter, under first party and strict liability respectively, with fully informed consumers and courts. Under first party liability, the marginal benefit of additional care (*FF*) is simply the reduction in probability of injury to the patient, which is assumed to be subject to diminishing returns. Under strict lia-

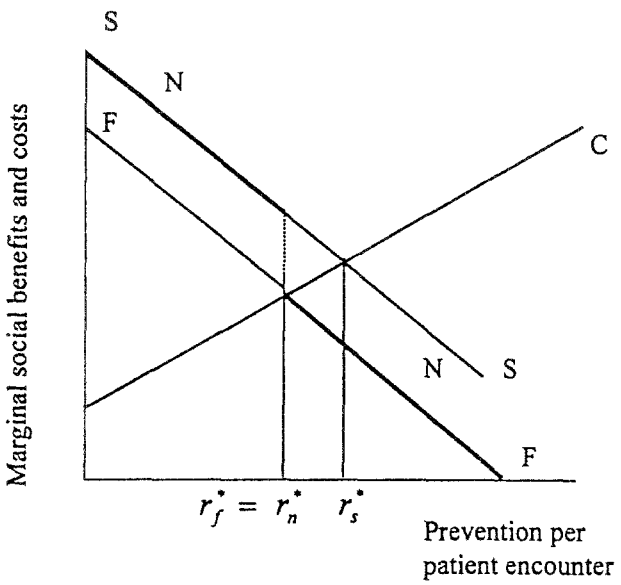

Figure 1. Optimal prevention under first party, negligence and strict liability.

bility, there is an additional benefit from care if the physician incurs some uninsurable loss of time, reputation or inconvenience in responding to claims, so $SS > FF$.

The curve labeled *CC* shows the marginal cost of care. The discontinuous heavy line *NN* shows the marginal social benefits of care under a negligence rule with the due care standard, r_n^*, optimally set equal to r_f^*, the level that would be chosen by fully informed patients. *NN* is discontinuous at r_n^* because for levels of care that equal or exceed that level, the physician meets the due care standard; hence he bears no liability.

Type 1 errors (failure to file or find liability in all instances of negligence) or a sub-optimal penalty if negligent imply a downward shifting of *NN*. Provided the vertical segment of *NN* intersects *CC* at r_n^*, physicians will still choose to be non-negligent. But if the Type 1 errors are sufficiently large, the intersection occurs to the left of r_n^*. It is then cheaper for physicians to practice with less than due care and to purchase liability insurance than to be non-negligent.

Type 2 errors consist of setting the standard to the right of r_n^*. As long as $r_f^* < r_n^* < r_s^*$, physicians will meet the standard. But if $r_n^* > r_s^*$, it is cheaper for physicians to choose r_s^*, i.e., to practice below the excessively high standard set by the courts and to insure against the resulting claims. Patients are thereby better off than if physicians adhered to the excessively high standard, but are worse off than if the standard were set at r_n^*, because medical fees rise to cover the additional prevention costs $(r_s^* - r_n^*)$ and physician's uninsured disutility of suit.[29]

Casual evidence suggests that Type 1 errors dominate the demand for liability insurance. In medical and other lines of professional liability, the courts defer to the customary practice of professionals "in good standing" as the standard of due care, rather than apply the Hand cost-benefit calculus in each case. This creates a bulwark against Type 2 errors for above-average physicians but a presumption of Type 2 errors and consequent incentive to insure for those of below-average competence. Thus, the Type 2 errors hypothesis might predict that a substantial fraction of physicians would not buy insurance, which is not the case. On the other hand, Type 1 errors, which give all physicians an incentive to buy insurance, are very common (see Section 3.2).

References

Abraham, K.S. and P.C. Weiler (1994), "Enterprise medical liability and the evolution of the American health system", Harvard Law Review 108:381–438.

ACS (1996), What is Medical Misadventure (Accident Rehabilitation and Compensation Insurance Corporation, Wellington, New Zealand).

Bovbjerg, R.R. (1995), Medical Malpractice: Problems and Reforms (The Urban Institute, Washington, DC).

Bovbjerg, R.R., F.A. Sloan, et al. (1989), "Valuing life and limb in tort: Scheduling pain and suffering", Northwestern University Law Review 83(4):908–976.

Brown, J.P. (1973), "Toward an economic theory of liability", Journal of Legal Studies 2:323–349.

[29] In principle, the additional cost of liability insurance is offset by a reduction in first party insurance costs.

Butler, R.J., and J.D. Worral (1983), "Workers' compensation: Benefits and injury claim rates in the 1970s", Review of Economics and Statistics 65:580–589.
Coase, R. (1960), "The problem of social cost", Journal of Law and Economics 3:1–44.
Cook, P.J., and D.A. Graham (1977), "The demand for insurance and protection: The case of irreplaceable commodities", Quarterly Journal of Economics 91:143–156.
Craswell, R., and J. Calfee (1986), "Deterrence and uncertain legal standards", Journal of Law, Economics and Organization 2(2):279–303.
Cummins, J.D., and P.M. Danzon (1997), "Quality, price and capital flows in insurance markets", Journal of Financial Intermediation 6(1):3–38.
Cummins, J.D., and J.F. Outreville (1987), "An international analysis of underwriting cycles in property liability insurance", Journal of Risk and Insurance 54(2):246–262.
Danzon, P.M. (1983), "Contingent fees for personal injury litigation", Bell Journal of Economics 14:213–230.
Danzon, P.M. (1984a), "The frequency and severity of medical malpractice claims", Journal of Law and Economics 27:115–148.
Danzon, P.M. (1984b), "Tort reform and the role of government in private insurance markets", Journal of Legal Studies 13(3):517–549.
Danzon, P.M. (1985a), Medical Malpractice: Theory, Evidence and Public Policy (Harvard University Press, Cambridge).
Danzon, P.M. (1985b), "Liability and liability insurance for medical malpractice", Journal of Health Economics 4:309–331.
Danzon, P.M. (1986), "New evidence on the frequency and severity of medical malpractice claims", Law and Contemporary Problems 5(49):57–84.
Danzon, P.M. (1990a), "The "crisis" in medical malpractice: A comparison of trends in the United States, Canada, the United Kingdom and Australia", Law, Medicine and Health Care 18:48–58.
Danzon, P.M. (1990b), Medical Malpractice: Incidence and Incentive Effects (The Wharton School, University of Pennsylvania, Philadelphia).
Danzon, P.M. (1991), "Liability for medical malpractice", Journal of Economic Perspectives 5:51–69.
Danzon, P.M. (1994a), "The Swedish patient compensation system: Myths and realities", International Review of Law and Economics 14:453–466.
Danzon, P.M. (1994b), "Alternative liability regimes for medical malpractice: Results from simulation analysis", Journal of Risk and Insurance 61(2):219–244.
Danzon, P.M. (1994d), "Tort reform: The case of medical malpractice", Oxford Review of Economic Policy 10(1):84–98.
Danzon, P.M. (1997), "Tort liability: A minefield for managed care", Journal of Legal Studies 26(2):491–519.
Danzon, P.M., and L. Lillard (1983), "Settlement out of court: The disposition of medical malpractice claims", Journal of Legal Studies 12:345–377.
Danzon, P.M., M.V. Pauly and R. Kington (1990), "The effects of malpractice litigation on physicians' fees and incomes", American Economic Review 80(2):122–127.
Dewees, D., P. Coyte, et al. (1989), Canadian Medical Malpractice Liability: An Empirical Analysis of Recent Trends (University of Toronto).
Dobson, F. (1998), Dobson to Tackle Rising Levels of Litigation in the Health Service (Department of Health, London).
Doherty, N.A., and G. Dionne (1993), "Insurance with undiversifiable risk: Contact structure and organizational form of insurance firms", Journal of Risk and Uncertainty 6:187–203.
Eddy, D.M. (1998), "Performance measurement: Problems and solutions", Health Affairs 17(4):7–25.
Ehrlich, I., and G.S. Becker (1972), "Market insurance, self-insurance and self-protection", Journal of Political Economy 80:623–648.
Ellis, R.P., C.L. Gallup, et al. (1990), "Should medical professional liability insurance be experience rated?" Journal of Risk and Insurance 57(1):66–78.
Entman, S., C. Glass, et al. (1994), "The relationship between malpractice claims history and subsequent obstetric care", Journal of the American Medical Association 272(20):1588–1591.

Epstein, R. (1978), "Medical malpractice: Its cause and cure", in: S. Rottenberg, ed., The Economics of Medical Malpractice (American Enterprise Institute Press, Washington, DC).

Farber, H.S., and M.J. White (1991), "Medical malpractice: An empirical examination of the litigation process", Rand Journal of Economics 22(2):199–217.

Farley, P.J. (1986), "Theories of the price and quantity of physicians' services: A synthesis and critique", Journal of Health Economics 5:315–333.

Fenn, P. (1993), "Compensation for medical injury: A review of policy options", in: C. Vincent, M. Ennis and B. Audley, eds., Medical Accidents (Oxford University Press, Oxford).

General Accounting Office (GAO) (1994), Medical Malpractice: Maine's Use of Practice Guidelines to Reduce Costs (US Government Printing Office, Washington, DC).

Ham, C., R. Dingwall, et al. (1988), Medical Negligence: Compensation and Accountability (The Kings Fund, London).

Ham, C., and S. Pickard (1998), Tragic Choices in Health Care: the Case of Child B (The Kings Fund, London).

Harrington, S., and P.M. Danzon (1994), "Price cutting in liability insurance markets", Journal of Business 67(4):511–538.

Harrington, S., and R. Litan (1988), "Causes of the Liability Insurance Crisis", Science 239:737–741.

Havighurst, C.C. (1995), Health Care Choices: Private Contracts as Instruments of Health Reform (American Enterprise Institute Press, Washington, DC).

Havighurst, C.C. (1997), "Making health plans accountable for the quality of care", Georgia Law Review 31(2):587–647.

Henderson, J.A., and T. Eisenberg (1988), "The quiet revolution in products liability: An empirical study of legal change", UCLA Law Review 37:479–553.

Hickson, G., E. Clayton, et al. (1994), "Obstetricians' prior malpractice experience and patients' satisfaction with care", Journal of the American Medical Association 272(20):1583–1587.

Kessler, D., and M. McClellan (1996), "Do doctors practice defensive medicine?", Quarterly Journal of Economics 111(2):353–390.

Localio, A.R., A.G. Lawthers, et al. (1993), "Relationship between malpractice claims and caesarian delivery", Journal of the American Medical Association 269(3):366–373.

McGuire, T.G. (2000), "Physician agency", in: A.J. Culyer and J.P. Newhouse, Handbook of Health Economics (Elsevier, Amsterdam) Chapter 9.

Mills, D.H., J.S. Boyden, et al. (1977), Report on the Medical Insurance Feasibility Study (California Medical Association and California Hospital Association, Sutter Publications, San Francisco, CA).

Nye, B.F., and A.E. Hofflander (1987), "Economics of oligopoly: Medical malpractice insurance as a classic illustration", Journal of Risk and Insurance 54(3):502–519.

NZBR (1998), Accident Compensation: Options for Reform (New Zealand Business Roundtable, Wellington, New Zealand).

Pauly, M.V. (1968), "The economics of moral hazard: comment", American Economic Review 58:531–537.

Pauly, M.V. (1980), Doctors and Their Workshops (University of Chicago Press, Chicago, IL).

Phelps, C.E. (2000), "Information diffusion and best practice adoption", in: A.J. Culyer and J.P. Newhouse, Handbook of Health Economics (Elsevier, Amsterdam) Chapter 5.

Posner, R.A. (1972), "A theory of negligence", Journal of Legal Studies 1:28–96.

Priest, G. (1987), "The current insurance crisis and modern tort law", Yale Law Journal 96(7):1521–1590.

Rea, S. (1981), "Lump sum vs. periodic damage awards", Journal of Legal Studies 10:131–154.

Reynolds, R.A., J.A. Rizzo, et al. (1987), "The cost of medical professional liability", Journal of the American Medical Association 257(20):2776–2781.

Rolph, J.E. (1981), "Some Statistical Evidence on Merit Rating in Medical Malpractice Insurance", Journal of Risk and Insurance 48:247–260.

Rolph, J.E., R.L. Kravitz, et al. (1991), "Malpractice claims data as a quality improvement tool: Is targetting effective?", Journal of the American Medical Association 266(15):2093–2097.

Rosenberg, J.D., and H.J. Folberg (1994), "Alternative dispute resolution in a civil justice reform act demonstration district: Findings, implications and recommendations", Stanford Law Review 46:1487–1551.

Rubin, R.J., and D.N. Mendelson (1993), Defensive Medicine and Medical Liability Reform: Estimating Costs and Potential Savings (Lewin-VHI, Fairfax, VA).

Schwartz, W.B., and D.N. Mendelson (1989), "The role of physician-owned insurance companies in the detection and deterrence of negligence", Journal of the American Medical Association 262(10):1342–1346.

Schwartz, W.B., and D.N. Mendelson (1989), "Physicians who have lost their malpractice insurance: Their demographic characteristics and the surplus lines companies that insure them", Journal of the American Medical Association 262(10):1335–1341.

Shanley, M.G., and M.A. Peterson (1987), Posttrial Adjustments to Jury Awards (The RAND Corporation, Santa Monica, CA).

Shavell, S. (1980), "Strict liability vs. negligence", Journal of Legal Studies 9:1–25.

Shavell, S. (1982), "On liability and insurance", Bell Journal of Economics 13:120–132.

Shavell, S. (1997), "The fundamental divergence between the private and the social motive to use the legal system", Journal of Legal Studies 26(2, Pt. 2):575–612.

Sloan, F.A. (1998), "Automobile accidents", in: P. Newman, ed., The New Palgrave Dictionary of Economics and the Law (Macmillan, London).

Sloan, F.A., and T. Hoerger (1991), "Uncertainty, information and the resolution of malpractice disputes", Journal of Risk and Uncertainty 4(4):343–363.

Sloan, F.A., P.B. Githens, et al. (1993), Suing for Medical Malpractice (University of Chicago Press, Chicago, IL).

Sloan, F.A., P.M. Mergenhagen, et al. (1989), "Medical malpractice experience of physicians: Predictable or haphazard?", Journal of the American Medical Association 262(23):3291–3297.

Smith, R. (1990), "The epidemiology of malpractice: Adverse events common, negligent cases rare", British Medical Journal 301:621.

Spence, M. (1977), "Consumer misperceptions, product failure and product liability", Review of Economic Studies 64:561–572.

Towse, A., and P.M. Danzon (1998), An Economic Analysis Of Clinical Negligence: The Options For The NHS (Office of Health Economics, London).

Weiler, P.C. (1991), Medical Malpractice on Trial (Harvard University Press, Cambridge).

Weiler, P.C., H.H. Hiatt, et al. (1993), A Measure of Malpractice: Medical Injury, Malpractice Litigation and Patient Compensation (Harvard University Press, Cambridge).

White, M.J. (1994), "The value of liability in medical malpractice", Health Affairs 13(4):75–87.

Winter, R.A. (1988), "The liability crisis and the dynamics of competitive insurance markets", Yale Journal of Regulation 5(2):455–500.

Zuckerman, S., R. Bovbjerg, et al. (1990), "Effects of tort reforms and other factors on medical malpractice insurance premiums", Inquiry 27(2):167–182.