

EFFICIENCY GAINS FROM HEALTH INSURANCE

Joseph G. Eisenhauer, Wright State University

ABSTRACT

For an individual requiring medical care, the initial units of treatment are likely to be valued above their production cost, while excessive units are undervalued. It is well-known that by reducing the effective price of care to the patient, health insurance encourages the over-utilization of treatment, resulting in inefficiency. However, by transferring resources to the claimant, health insurance also raises the patient's demand for medical care, increasing the proportion of care which is efficient. For a sufficiently strong response to the income transfer, the net result of health insurance is an increase in economic efficiency. *JEL classifications:* G22, I11.

INTRODUCTION

It is well known that the over-utilization of medical care induced by health insurance causes inefficiency. With insurance covering most of their medical expenses, patients receive treatment that they value below production cost and would not otherwise purchase (Pauly, 1968). It is less widely appreciated that insurance also raises a claimant's income, which increases the demand for medical care and diminishes the deadweight loss (de Meza, 1983). The present note extends that analysis by showing that a sufficiently strong response to the income transfer can more than offset the efficiency loss and result in a net efficiency gain.

The first section below discusses the economic theory behind this result. The second section develops a simple mathematical model and a graphical depiction, while the third section provides a numerical example to illustrate the theory. The fourth and fifth sections examine deadweight gains and losses under various parameterizations of the model. The conclusion provides some caveats and policy implications.

ECONOMIC THEORY

A basic principle of consumer theory is diminishing marginal utility: as more of a good is consumed, *ceteris paribus*, each additional unit is worth less to the consumer than the previous unit. This premise underlies the usual downward-sloping shape of a demand curve, reflecting the inverse relationship between the per-unit price that an individual is willing to pay for a good (or service) and the quantity consumed. This notion is as applicable to health care as to any other good or service. For an ailing individual, the first few units of treatment are likely to matter the most—indeed, they may be life-saving—and will therefore carry the greatest personal value. To the extent that a patient would be willing to pay more for each such unit of treatment than its market price, the individual receives a windfall measured by consumer surplus: the difference between one's willingness-to-pay and one's actual payment on the initial units. Because the individual's willingness-to-pay declines

with consumption while the cost of production does not, economic theory states that in the absence of third-party payments, an ill individual should continue to purchase medical care just up to the point at which the value (s)he places on the final unit (or fraction thereof) exactly equals its market price. If more than that quantity is consumed, inefficiency arises because the value of the additional care to the individual falls below its cost of production.

Applying this basic idea, Pauly (1968) considered how health insurance affects the quantity of medical care demanded. Because an insurance policy specifies a fractional rate of copayment, the insured effectively faces a reduced price per unit, and upon becoming ill opts for more treatment than (s)he would at full cost. The excess utilization of medical care represents *ex post* moral hazard.¹ The inefficient allocation of treatment is measured in Pauly's model by deadweight loss, or the difference between the cost of production and the individual's willingness-to-pay for the excess units. That model became the basis for estimates of sizable deadweight losses (Pauly, 1969; Feldstein, 1973; Feldstein and Friedman, 1977), and provided an intellectual foundation for the managed care movement, in which insurers more carefully scrutinized the medical treatment provided to claimants.

But economic theory also posits that demand curves shift in response to changes in income. For normal goods, including medicine, the income-elasticity of demand is positive, implying an outward (inward) shift of the demand curve when income rises (falls). This aspect of the demand for medical care was overlooked until de Meza (1983) observed that an insurance claimant receives the equivalent of an income transfer from the insurance pool. Provided that the claim exceeds the insurance premium, the net increase in income to the claimant shifts the demand curve for medical care outward, reducing the deadweight loss and increasing the consumer surplus. For a sufficiently strong response, the overall effect of insurance may be an increase in economic efficiency, as demonstrated below.

MODEL

Consider an individual whose demand for medical care in a state of illness is given by

$$Q = (\alpha Y^\gamma - P) / \beta \quad (1)$$

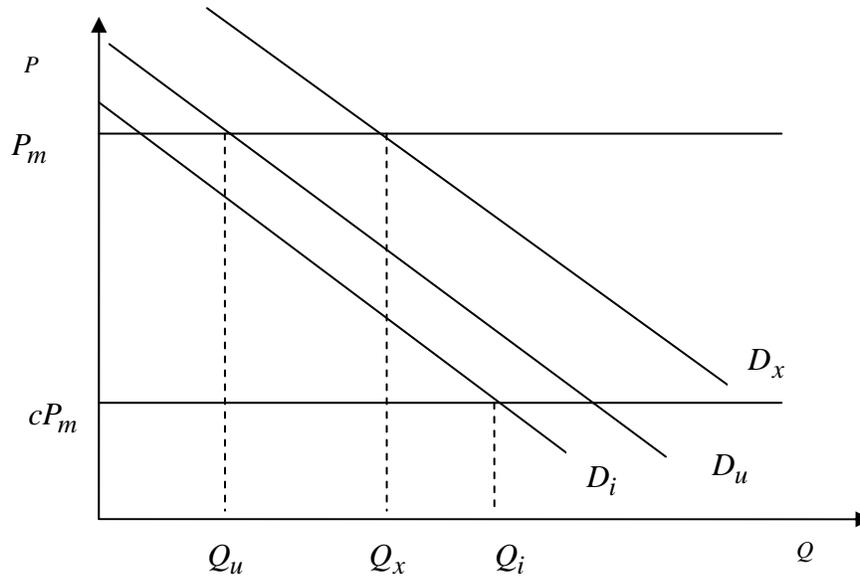
where Q denotes the quantity of care, P is the per unit price of care to the patient, and Y is the individual's disposable income. This particular functional form is chosen for convenience because it is linear in P and Q and allows parallel shifts of the demand curve in response to changes in income.² As shown below, the parameters α and γ are related to the price- and income-elasticities of demand, and β is a scaling factor. To illustrate the model, we follow the prior literature by assuming that each unit of medical care (for example, a day of hospitalization) is sold at a market price of P_m , which is equal to its fixed production cost.

Given this demand function, an individual who becomes ill without health insurance would purchase

$$Q_u = (\alpha Y^\gamma - P_m) / \beta \quad (2)$$

units of medical care, where the subscript u identifies the individual as uninsured. This result is depicted in Figure 1 at the intersection of the demand curve D_u and the market price P_m .

FIGURE 1
SHIFTS OF THE DEMAND CURVE FOR MEDICAL CARE
IN RESPONSE TO INCOME CHANGES



Now suppose that, prior to becoming ill, the same individual had purchased a health insurance policy with a coinsurance rate of c at a premium of π . Then the effective price of medical care to the patient is cP_m per unit, and his or her available income is $Y - \pi$. Under these conditions, the individual chooses to consume

$$Q_i = [\alpha(Y - \pi)^\gamma - cP_m] / \beta \quad (3)$$

units of care, where the subscript i indicates that the individual is privately insured. Figure 1 illustrates the increase in treatment. The payment of the insurance premium (i.e., the income reduction) causes the demand curve for medical care in a state of illness to shift inward slightly, as depicted by the curve D_i . However, the reduction in the price faced by the patient (to cP_m per unit) encourages a movement along the new demand curve, to Q_i units of care. The total production cost of the care provided to the patient is now $P_m Q_i$. The additional $Q_i - Q_u$ units of care utilized as the result of insurance represent *ex post* moral hazard and have traditionally been

viewed as inefficient, precisely because the individual would not have purchased this much medical treatment without health insurance.

However, by paying claims equal to the fraction $1 - c$ of the insured's medical expenses, the insurer has, in effect, provided an earmarked subsidy or in-kind transfer to the claimant. To evaluate its efficiency, de Meza (1983) considers how this insurance transfer would have been spent if it had been paid to the claimant in cash, rather than reimbursing the expenditure on medical care. In particular, imagine that upon becoming ill, the insured had obtained a diagnosis and an estimated medical bill of $P_m Q_i$ from a physician, had filed a claim with the insurance company, and had received a check equal to the fraction $1 - c$ of the total bill, which (s)he could then cash. Because medical care is a normal good, some portion of the additional income transferred to the claimant would still have been spent on medical care, while the rest would have been diverted to the purchase of other goods. Clearly, the former is consistent with the fundamental purpose of insurance, and the latter is not. Thus, some initial fraction of the in-kind transfer that actually occurs is used efficiently. Only the latter portion of the transfer is truly inefficient, representing an expenditure on medical care that the individual would rather not have made.

The efficient and inefficient components can be distinguished by extending the model as follows. Because the individual who paid a premium of π and received a cash claim of $(1 - c)P_m Q_i$ would face the full market price per unit, the quantity of care demanded would be

$$Q_x = [\alpha(Y - \pi + (1 - c)P_m Q_i)^\gamma - P_m] / \beta \quad (4)$$

units, where the subscript x denotes the presence of a (hypothetical) cash transfer. In Figure 1, the cash claim shifts the patient's demand curve for medical treatment outward from D_i to D_x where (s)he consumes Q_x units at the market price. Although the total moral hazard is measured by $Q_i - Q_u$, the units from Q_u to Q_x are an efficient use of insurance, because this extra treatment represents care that the insured values at or above its market cost, when given the income transfer. Only the final $Q_i - Q_x$ units of care that are utilized by the insured in practice (i.e., under an in-kind transfer) represent inefficiency. Below, we demonstrate the potential for a net efficiency gain.

ILLUSTRATION

To illustrate the model, let us initially assume that each unit of care has a fixed market price of $P_m = \$2000$, and that the patient is endowed with a disposable income of \$50,000. From the demand function (1), the price-elasticity of demand for medical care is $\eta = (\partial Q / \partial P)(P / Q) = -P / (\alpha Y^\gamma - P)$ and the income-elasticity of demand is given by $\varepsilon = (\partial Q / \partial Y)(Y / Q) = \alpha \gamma Y^\gamma / (\alpha Y^\gamma - P)$. Empirical research has typically found price-elasticities in the neighborhood of -0.25 and income-elasticities near 0.30; see for example the surveys by Zweifel and Manning (2000) and Ringel et al. (2002). Inserting these elasticities into the model

at $P = P_m = 2000$ implies $\gamma = .24$ and $\alpha = 745$.³ The scaling factor, β , is initially set at 500; after the initial analysis, we will examine the effects of alternative parameter values.

Given the initial calibration, equation (2) indicates that an uninsured patient purchases $Q_u = 16$ units of care. However, if (s)he had purchased a health insurance policy with a 30 percent coinsurance rate for a premium of \$1500, his or her available income would be \$48,500 and the effective price per unit of medical care would become \$600. With such insurance, equation (3) shows that the individual would consume $Q_i = 18.65$ units of treatment when ill.⁴ Note that the total production cost of the care provided to the individual would now be $\$2000 \times 18.65 = \$37,300$, which would have been an affordable but not optimal expenditure for this individual in the absence of insurance.⁵ The difference between Q_i and Q_u represents 2.65 total units of *ex post* moral hazard.

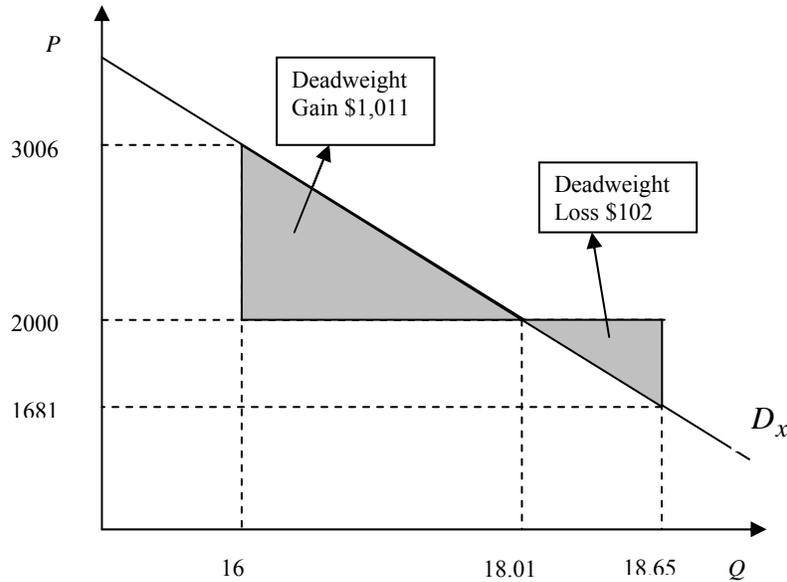
In this example, the insured spends \$11,190 on copayments and files claims of \$26,110 after having paid a premium of \$1,500 for the insurance policy. The insured has therefore received a net transfer worth $\$26,110 - \$1,500 = \$24,610$ from the insurance pool. If the claim had been paid in cash, the claimant would have had $\$50,000 + 24,610 = \$74,610$ in available income to spend on medical care and other goods. At the full market price, the quantity of care demanded would have been 18.01 units, as per equation (4). Thus, in the present example, 2.01 out of 2.65 units, or about 75.85 percent of the additional care induced by insurance is efficient, and 24.15 percent is inefficient.⁶

DEADWEIGHT LOSS AND DEADWEIGHT GAIN

To appreciate the economic significance of the transfer effect, however, it is necessary to translate the outcome obtained above into measures of deadweight loss and deadweight gain. Because the demand curve denoted D_x is linear, it is a simple matter to measure the deadweight loss and deadweight gain by the shaded triangular areas depicted in Figure 2. Along this demand curve, the individual under-values each unit of care (or fraction thereof) beyond 18.01 units; thus, the entire 18.65 units of treatment (an amount equal to Q_i) would only be purchased with cash if the price per unit was \$1,681 rather than \$2,000.⁷ The deadweight loss is therefore calculated as $.5(\$2,000 - \$1,681)(18.65 - 18.01) = \102 , which represents the extent to which the final 0.64 units are under-valued by the insured.

Yet along this same demand curve, the *first* 18.01 units of treatment are valued *above* the production cost of \$2,000 per unit, so the individual gains consumer surplus from them. Because the individual would be willing to pay up to \$3,006 for the 16th unit, the efficient moral hazard—the extra 2.01 units utilized as the result of insurance—generates an increase in consumer surplus (or deadweight gain) equal to $.5(\$3,006 - \$2,000)(18.01 - 16) = \$1,011$.⁸ This welfare gain results from the fact that the effective transfer of income to the insured has shifted the demand curve outward, raising the individual's willingness-to-pay for medical care, so that each fraction of a unit between 16 and 18.01 is now worth more to the insured than its production cost. The net result is therefore an *efficiency gain* of $\$1,011 - \$102 = \$909$, or about 2.4 percent of the \$37,300 production cost of providing care to the insured.

FIGURE 2
EXAMPLE OF DEADWEIGHT GAIN AND DEADWEIGHT LOSS



Indeed, because the demand function used here is linear in P and Q , there will be a net efficiency gain whenever more than half of the moral hazard is efficient; that is, whenever $Q_x - Q_u > Q_i - Q_x$. Specifically, the ratio of the deadweight gain to the deadweight loss is equal to the square of the ratio of efficient moral hazard to inefficient moral hazard:⁹

$$\frac{Gain}{Loss} = \left(\frac{Q_x - Q_u}{Q_i - Q_x} \right)^2. \quad (5)$$

Thus, with about 75.85 percent of moral hazard being efficient, the efficiency gain/loss ratio in this scenario is roughly $(0.7585/0.2415)^2 = 9.9$. The next section recalibrates the model for different values of the parameters.

ALTERNATIVE PARAMETERIZATIONS

The first row of Table 1 presents the efficiency results for the base case, and the subsequent rows provide several additional examples based on changes in the parameter values. Because the inefficiency of health insurance arises as a result of price-elasticity in the demand for medical care, the efficiency gains are smaller when price effects are greater. Conversely, because the increase in consumer surplus is due to the effect of the insurance transfer on the claimant's income, the efficiency gains are stronger when there is greater income-elasticity in the demand curve.¹⁰ In the

present context, efficiency gains are greatest with relatively high α , low β , high γ , low c , high P_m , and a low premium for the insurance policy, *ceteris paribus*. And as the table indicates, efficiency gains are also stronger when the initial value of disposable income (Y) is lower than when it is higher.

TABLE 1.
EFFICIENCY RESULTS FOR SELECTED PARAMETER VALUES*

α	β	γ	c	P_m	Y	Premium	Q_u	Q_i	Q_x	Dead weight Gain \$	Dead weight Loss \$	Net Gain \$	Net Gain %
745	500	.24	.30	2000	50,000	1500	16.00	18.65	18.01	1011.03	102.08	908.95	2.44
845	500	.24	.30	2000	50,000	1500	18.68	21.31	21.26	1665.09	0.73	1664.36	3.90
645	500	.24	.30	2000	50,000	1500	13.31	15.99	14.82	571.79	337.31	234.48	0.73
745	600	.24	.30	2000	50,000	1500	13.33	15.54	14.75	602.72	189.39	413.33	1.33
745	400	.24	.30	2000	50,000	1500	19.99	23.31	23.07	1896.78	11.36	1885.42	4.04
745	500	.25	.30	2000	50,000	1500	18.28	20.91	20.88	1692.65	0.21	1692.44	4.05
745	500	.23	.30	2000	50,000	1500	13.95	16.62	15.50	604.86	313.19	291.68	0.88
745	500	.24	.35	2000	50,000	1500	16.00	18.45	17.86	868.61	87.14	781.47	2.12
745	500	.24	.25	2000	50,000	1500	16.00	18.85	18.16	1174.77	117.86	1056.91	2.80
745	500	.24	.30	2500	50,000	1500	15.00	18.35	17.42	1474.66	214.21	1260.45	2.75
745	500	.24	.30	1500	50,000	1500	17.00	18.95	18.56	609.63	38.57	571.06	2.01
745	500	.24	.30	2000	70,000	1500	17.68	20.37	19.44	779.71	212.32	567.39	1.39
745	500	.24	.30	2000	30,000	1500	13.69	16.27	16.12	1474.91	5.98	1468.94	4.51
745	500	.24	.30	2000	50,000	1700	16.00	18.63	18.00	999.97	100.71	899.26	2.41
745	500	.24	.30	2000	50,000	1300	16.00	18.70	18.03	1032.45	102.97	929.48	2.49

* The first row represents the base case described in the text; alternative parameter values are highlighted in subsequent rows.

Of course, it would be an equally simple matter to derive numerical examples in which the deadweight loss exceeds the increase in consumer surplus so that the net result of insurance is inefficiency. In particular, such an outcome is obtained when the demand for medical care exhibits relatively strong price-elasticity and limited income-elasticity. Further research is therefore necessary to establish the prevalence of gains and losses in practice. However, the model above provides a useful framework for analysis, and it is clear from the numerical examples that under empirically plausible parameter values, health insurance is capable of generating efficiency improvements rather than allocative inefficiencies.

CONCLUSION

Health insurance operates by pooling the premiums of policyholders to pay claims for medical care. Claims therefore transfer income from policyholders who remain healthy to those unfortunate few who fall ill in any given period. To the patient, the first units of medical care are worth more than they cost to produce. Thus, up to some point, the funds transferred through the claims mechanism to buy medical treatment are used efficiently. Beyond that point, the value of extra medical care to the patient falls below its production cost, creating inefficiency. Consequently, the net result of health insurance may be either efficiency gains or losses. The model developed above illustrates the real possibility that efficiency gains are dominant.

While the present analysis has been framed in terms of privately purchased insurance, the same model can be applied to social insurance programs such as Medicare, Medicaid, or the State Children's Health Insurance Program (SCHIP), which are financed through taxation. Indeed, the original analysis by Pauly (1968) was undertaken in response to Arrow's (1963) classic paper which argued that governments should provide health insurance whenever private markets fail to do so. Much of the subsequent controversy over the ensuing four decades has focused on the relative efficiency of public health insurance programs. Recently, both the expansion of Medicaid and the continuation of the SCHIP program have been hotly contested, and several state legislatures have initiated their own public programs.¹¹ The current paper contributes to this debate by demonstrating the potential for achieving efficiency gains through health insurance, whether public or private.¹² In this respect, perhaps the most interesting result is the finding that efficiency gains from health insurance are largest when initial income is low, which suggests that providing coverage to the poor is most likely to generate efficiency gains.

A number of refinements and extensions of this simple model are possible. In defining Y to be disposable income, for example, the role of taxation has been oversimplified. Eisenhower (2002) points out that under current tax law, both health insurance premiums and uninsured medical expenses exceeding 7.5 percent of adjusted gross income are tax deductible. The tax deduction for uninsured losses provides a form of public insurance which effectively reduces the demand for private insurance and may simultaneously encourage the uninsured to utilize more treatment than they otherwise would. If these effects were incorporated into the present model, the difference in utilization of medical care between the insured and the uninsured—i.e., the extent of the moral hazard—would be diminished. Deductibility of uninsured losses, however, is a somewhat arbitrary feature of the current income tax code rather than an inherent characteristic of health insurance, and for this reason has not been addressed in the present work.

Moreover, the present analysis contrasts the utilization of medical care in the absence of insurance with the treatment provided under a particular health insurance policy. A more subtle question involves the efficiency differences among various types of health insurance policies. Many health insurance plans, both employer-sponsored and individually-purchased, allow policyholders some choice. For example, more extensive coverage (i.e., more insured ailments, lower deductibles, and/or lower copayments) or greater flexibility (a wider selection of providers) may be available with higher premiums. Which of several such policy options would generate the greatest efficiency (or least inefficiency) is itself an interesting topic for future research.¹³

Perhaps more importantly, the specification of the demand function for medical care involves relatively strong assumptions. Although the price- and income-elasticities adopted here are within the empirically observable range, it is not clear that demand is necessarily either linear or subject to parallel shifts in response to income changes; these simplifying assumptions should therefore be relaxed in future work.

REFERENCES

- Arrow, Kenneth J. 1963. "Uncertainty and the Welfare Effects of Medical Care." *American Economic Review* 53: 941-973.

- de Meza, David. 1983. "Health Insurance and the Demand for Medical Care." *Journal of Health Economics* 2: 47-54.
- Eisenhauer, Joseph G. 2002. "Relative Effects of Premium Loading and Tax Deductions on the Demand for Insurance." *Journal of Insurance Issues* 25: 7-62.
- Eisenhauer, Joseph G. 2006. "Severity of Illness and the Welfare Effects of Moral Hazard." *International Journal of Health Care Finance and Economics* 6: 290-299.
- Feldstein, Martin S. 1973. "The Welfare Loss of Excess Health Insurance." *Journal of Political Economy* 81(2, part 1): 251-280.
- Feldstein, Martin S. and Bernard Friedman. 1977. "Tax Subsidies, the Rational Demand for Health Insurance, and the Health Care Crisis." *Journal of Public Economics* 7: 155-178.
- Kaiser Commission on Medicaid and the Uninsured. 2007. *SCHIP Reauthorization: Key Questions in the Debate*, Washington, D.C.: Henry J. Kaiser Foundation, August 29.
- Muenning, Peter, Peter Franks, and Marthe Gold. 2005. "The Cost Effectiveness of Health Insurance." *American Journal of Preventive Medicine* 28 (1): 59-64.
- National Conference of State Legislatures. 2007. "2007 Bills on Universal Health Care Coverage: Legislatures Fill in the Gaps." Washington, D.C.: National Conference of State Legislatures.
- Pauly, Mark V. 1968. "The Economics of Moral Hazard: Comment." *American Economic Review* 58: 531-537.
- Pauly, Mark V. 1969, "A Measure of the Welfare Cost of Moral Hazard." *Health Services Research* 4 (4): 281-292.
- Ringel, Jeanne S., Susan D. Hosek, Ben A. Vollaard, and Sergej Mahnovski. 2002. *The Elasticity of Demand for Health Care: A Review of the Literature and Its Application to the Military Health System*. Santa Monica, CA: RAND Corp.
- Santerre, Rexford E. 2006. "Examining the Marginal Access Value of Private Health Insurance." *Risk Management and Insurance Review* 9: 53-62.
- Zweifel, Peter and Willard G. Manning. 2000. "Moral Hazard and Consumer Incentives in Health Care." in Anthony J. Culyer and Joseph P. Newhouse (eds), *Handbook of Health Economics*, vol. 1A. Amsterdam: Elsevier.

ENDNOTES

1. Another potential source of inefficiency is *ex ante* moral hazard, which occurs if the insured takes fewer health precautions and thereby increases the likelihood of becoming ill.
2. Note that this demand curve can be written equivalently as $P = \alpha Y^\gamma - \beta Q$. It would be possible, of course, to complicate the model by allowing the demand curve of the ailing individual to change with different degrees of illness (see, for example, Eisenhauer, 2006), but for present purposes it will suffice to hold the severity constant when a state of illness occurs.
3. Algebraic manipulation yields $\gamma = \varepsilon / (1 - \eta) = 0.3 / 1.25 = 0.24$ and $\alpha = P(1 - \eta^{-1}) / Y^\gamma = 2000(1 + 4)50,000^{-.24} = 745.156$; to simplify, the latter is rounded to 745.

4. Of course, units of medical care may not be perfectly divisible as in the numerical example used here, but continuous demand functions are mathematically more convenient than functions requiring discrete units.
5. The extent to which insured care would be otherwise unaffordable to the patient has been called the access motive for buying health insurance; see for example, Santerre (2006). Because Q_i would be affordable without insurance, the access motive is not present in the current case.
6. The probability of illness can be recovered from information on premium loading and claims. Suppose, for example, that the policy has a loading factor of 1.5, implying that \$1,000 of the \$1,500 premium was pooled by the insurer and the other \$500 covered administrative expenses. Then the insured has contributed \$1,000 to the pool and filed claims for \$26,110. If this is typical, it implies that there are, on average, about 25 policyholders who do not file claims for each one that does—i.e., about 1 out of every 26 policyholders, or about 4 percent of all those insured, become ill and file claims during the period in question.
7. From the demand curve $P = 745Y^{.24} - 500Q$, an individual with \$74,610 in available income would consume 18.65 units only if $P = 1,681$.
8. From the demand curve $P = 745Y^{.24} - 500Q$, an individual with \$74,610 in available income would consume 16.00 units if $P = 3,006$.
9. To see this, note that along the demand curve D_x , $P_u = \alpha Y^\gamma - \beta Q_u$, $P_x = \alpha Y^\gamma - \beta Q_x$, and $P_i = \alpha Y^\gamma - \beta Q_i$. The deadweight gain is $.5(P_u - P_x)(Q_x - Q_u) = .5\beta(Q_x - Q_u)^2$ and the deadweight loss is $.5(P_x - P_i)(Q_i - Q_x) = .5\beta(Q_i - Q_x)^2$; their ratio gives equation (5).
10. For each of the cases in Table 1, the income-elasticity of demand for medical care lies between 0.28 and 0.32, while the price-elasticity of demand lies between -0.34 and -0.17; these values are consistent with the results of previous empirical research cited by Ringel et al. (2002).
11. See for example, the Kaiser Commission on Medicaid and the Uninsured (2007) and the National Conference of State Legislatures (2007).
12. Recent research has also demonstrated that public health insurance is efficient in a rather different sense. Muenning et al. (2005) show that the improvements in health and extension of life that are attributable to health insurance are as cost-effective as other public investments (such as airline safety) that society currently chooses to fund.
13. To some extent, simultaneous variations of c and π in the present model can accommodate such questions. For example, a reduction of the coinsurance rate from 30 percent to 29 percent, offset by a premium increase of \$370 (roughly one percent of production cost in the base case), *ceteris paribus*, would yield an efficiency result comparable to that of the base case presented above. Any number of other combinations of coinsurance rates and premiums could be compared in this manner.