To Wait or to Pay for Medical Treatment? Restraining ex-post Moral Hazard in Health Insurance

Stefan Felder
Die Berichte und weitere Informationen zu der Forschergruppe können abgerufen, werden unter:

The series „Priorisierung in der Medizin“ consists of working papers and research reports of the DFG (Deutsche Forschungsgemeinschaft, i.e., German Research Foundation) Research Group FOR655 „Priorisierung in der Medizin: eine theoretische und empirische Analyse unter besonderer Berücksichtigung der Gesetzlichen Krankenversicherung (GKV).(Prioritizing in Medicine: A Theoretical and Empirical Analysis in Consideration of the Public Health Insurance System)
Reports and further information can be found at
To Wait or to Pay for Medical Treatment? Restraining ex-post Moral Hazard in Health Insurance

Stefan Felder
Otto-von-Guericke University Magdeburg

We explore the hierarchy of two instruments, waiting time and coinsurance for medical treatment, for optimally solving the tradeoff between the economic gains from risk sharing and the losses from moral hazard. We show that the optimal waiting time is zero, given that the coinsurance rate is optimally set.

JEL Index: D82, H21, I18

Keywords: waiting time, coinsurance, moral hazard, optimal insurance contracts

1. Introduction

Most health insurance plans do not use indemnity payments that depend on health status, but reimburse the costs of the care actually provided to consumers. The reason is that the health status of a person can only be imperfectly observed. When insurance reimburses health care expenses, it subsidizes the price of health care at the margin, giving rise to a dead-weight loss. An optimal insurance contract thus involves copayments by the consumers. Coinsurance solves the tradeoff between risk sharing and the incentives to consume increased medical care.

---

Prof. Dr. Stefan Felder
Institute of Social Medicine and Health Economics (ISMHE)
Leipziger Str. 44
D-39120 Magdeburg
phone: 0391-6724321
e-mail: stefan.felder@ismhe.de
To Wait or to Pay for Medical Treatment?

User charges apply in social health insurance systems in many countries. Some services, such as dental care, may be excluded from the social insurance package, and deductibles and copayment rates are used to restrain the demand for health services. Countries with a national health service, however, do not generally employ coinsurance to curb demand for health care. They instead use waiting time for this purpose.

There is a vast literature analysing waiting times and waiting lists in the provision of health care services (for an overview see Cullis, Jones and Propper, 2000). Lindsay and Feigenbaum (1984) develop a model in which waiting list queues function as a rationing device. Waiting time matters because the value of health care decays the longer treatment is postponed after the diagnosis. Iverson (1993) outlines a supply side model in which waiting lists for inpatient care are the result of a political bargaining process over resources. Iverson (1997) and Olivella (2002) explore the effect of a private sector on waiting in the national health care system. Another strand of literature explores the scope of using waiting as a method of rationing health care in order to meet distributional objectives. Hoel and Sæther (2003) show that queuing in the public health care system might entice high-income persons to opt out and buy health care privately while still participating in the financing of the public system. If sufficient weight is given to equity, welfare can be improved by using the self-selection mechanism. In the absence of distributional motives, however, it is never optimal to have a positive waiting time for health treatment.

Smith (2005) analyzes user charges for different treatments under an exogenous health care budget. For a given treatment, all consumers have the same benefit, but, as the consumers’ wealth differs and marginal utility of wealth is decreasing, coinsurance reduces the demand for health services by discouraging poorer consumers from seeking treatment. In line with optimal taxation rules, the coinsurance rate for a treatment is inversely related to the price-elasticity of demand.

Gravelle and Siciliani (2006) extend the Smith model by introducing heterogeneity with respect to the consumers’ benefits from treatment and a public health care system which uses queuing to ration demand. Prioritization among treatments again follows an elasticity rule: the more elastic demand is with respect to the waiting time, the longer the optimal wait is. Adding user charges to the model, Gravelle and Siciliani demonstrate that Ramsey waits and prices apply simultaneously. However, they do not check whether the optimum requires the use of both instruments.

In this paper, abstracting from equity concerns, we address the complementary application of waiting time and coinsurance using a consumer’s health model, which includes health and consumption as arguments of the utility function and a linear coinsurance scheme designed to restrict moral hazard. We add the investment aspect of health, determining the consumer’s income in the process, and waiting time, which reduces income since an untreated ill person cannot work. We show that the optimal coinsurance rate is positive (due to moral hazard) and smaller than one (due to the consumer’s risk aversion). For a given coinsurance rate, the optimal waiting time may or may not be positive. However, if the coinsurance rate is optimally set, it is not optimal to have a positive waiting time for treatment.
2. The model

We consider a one-period world in which the representative consumer’s health status \( h \) in each state of the world \( s \) depends on a composite of health services, denoted by \( m \), and a random variable \( \theta > 0 \) which represents exogenous shocks to the consumer’s health status:

\[
h_s = m_s - \theta_s. \tag{1}
\]

The health status determines the consumer’s income in the labour market. We assume a production function with the following properties:

\[
f = f(h_s); f' > 0; f'' < 0; f''(0) = \infty. \tag{2}
\]

Waiting for treatment hinders the consumer from working and reduces her income by a factor \( \tau \), the waiting time, with \( 0 \leq \tau < 1 \). Since \( \theta \) is not observable to the insurer, the amount the insurer pays to the consumer in each state instead depends on her purchases of health services \( m \). Assume that the insurer charges a constant coinsurance rate \( \lambda \), with \( 0 \leq \lambda \leq 1 \). Moreover, we restrict the consumer’s total exposure to risk by assuming \( \lambda + \tau \leq 1 \).

With an actuarially fair insurance premium, the consumer’s consumption flow \( c \) in state \( s \) is restricted by the budget constraint:

\[
c_s = (1 - \tau) f(h_s) - \lambda m_s - (1 - \lambda) \sum_s \pi_s m_s, \tag{3}
\]

where units are chosen in such a way that all commodities sell at a price of unity and \( (1 - \lambda) \sum_s \pi_s m_s \) is the insurance premium with \( \pi \) as state-specific probabilities (\( \sum_s \pi_s = 1 \)).

The insurance premium may also represent a lump-sum tax to finance health insurance in a public system where the state is the only insurer.

The consumer’s utility depends on consumption \( c \) and health status \( h \). For a given value of \( \theta \) the consumer’s utility is given by

\[
u(\theta_s) = u(c_s, m_s - \theta_s). \tag{4}
\]

We assume

\[
u^c, u^m > 0; u^{cc}, u^{mm} < 0; u^{cm} = 0; u^c(0, m - \theta), u^m(c, 0) = \infty. \tag{5}
\]

The consumer faces a two-step optimization problem. After realization of \( \theta \), she decides on a utility maximizing consumption \( c \) and demand for health services \( m \).
with a given waiting time $\tau$ and coinsurance rate $\lambda$. Ex ante she solves for the optimal $\tau$ and $\lambda$ which maximize expected utility, taking into account her ex-post decision on $c$ and $m$.

The first-order condition for the ex-post maximum in state $s$ is

$$u^c(c^*_s, m^*_s - \theta) + u^c(c^*_s, m^*_s - \theta) f'(m^*_s - \theta_s)(1-\tau) = \lambda u^c(c^*_s, m^*_s - \theta_s).$$

Demand for health care is optimal when the marginal utility of health care equals its marginal cost. The marginal benefit from health care arises either from the higher utility of health or from a higher income, which is tantamount to the consumption model and the investment model of health care demand (Grossmann, 1972). The consumer’s net price of health care demand is $\lambda - f'(h^*_s)(1-\tau) > 0$.

From (6) and the budget constraint (3) with a fixed premium, we find:

$$\frac{\partial h^*_s}{\partial \theta} = \frac{-\lambda u^c(\lambda - f'(1-\tau))}{D} < 0,$$

$$\frac{\partial m^*_s}{\partial \theta} = \frac{(1-\lambda) u^c(\lambda - f'(1-\tau))}{D} > 0$$

$$\frac{\partial c^*_s}{\partial \theta} = \frac{-\lambda u^c(\lambda - f'(1-\tau))(2-\lambda - \tau)}{D} < 0,$$

with $D = \frac{u^{hc} + u^{cc}(\lambda - f'(1-\tau))^2 + (1-\tau)f^*u^c}{D} < 0$. The consumer will choose more health services when the shock to the health status is large. Demands for health and consumption will decrease and marginal utilities of health and consumption will in turn increase with a larger health shock.

Similarly, we derive Slutsky equations for the effect of a respective change in $\lambda$ and $\tau$ on the demand for health services:

$$\frac{\partial m^*_s}{\partial \lambda} = \frac{u^c + u^{cc} m^*_s (\lambda - f'(1-\tau))}{D} < 0,$$

$$\frac{\partial m^*_s}{\partial \tau} = \frac{u^c f'(1-\tau) + u^{cc} f(\lambda - f'(1-\tau))}{D} < 0.$$
To compare the relative sizes of the substitution and income effects, we calculate:

$$\frac{\partial m^*_s}{\partial \lambda} - \frac{\partial m^*_s}{\partial \tau} = \frac{u^c (1 - f') - u^c (\lambda - f'(1 - \tau))(m^*_s - f)}{D}.$$  \hspace{1cm} (12)

The substitution effect of an increase in coinsurance is larger than that of an increase in waiting time provided that $f' < 1$ (which follows from the first-order condition (6) given that $\lambda + \tau \leq 1$ and $u^c, u^m > 0$). When the health status has no effect on income (i.e. $f' = 0$), an increase in waiting time induces an income effect only. The income effect of coinsurance will be smaller when it covers a smaller base than the waiting time, i.e. $m^*_s < f$. A marginal increase in the coinsurance rate then involves a smaller monetary risk for the consumer than a marginal increase in the waiting time.

3. The optimal insurance contract

The optimal contract $(\lambda^*, \tau^*)$ maximizing the consumer’s expected (indirect) utility results from solving the program

$$\max_{\lambda, \tau} EV(\lambda, \tau) = \sum_s \pi_s u^c(c^*_s, m^*_s - \theta_s)$$  \hspace{1cm} (13)

subject to (3), $0 \leq \lambda \leq 1$ and $0 \leq \tau < 1$. For a marginal change in $\lambda$, the envelope theorem yields:

$$\frac{\partial EV}{\partial \lambda} = \sum_s \pi_s u^c(c^*_s, m^*_s - \theta_s) \left[-m^*_s + \sum_s \pi_s, m^*_s - (1 - \lambda) \sum_s \pi_s, \frac{\partial m^*_s}{\partial \lambda}\right].$$  \hspace{1cm} (14)

An increase in the coinsurance rate has two opposite effects on the consumer’s expected utility. On the one hand, it lowers expected utility since the consumer’s copayment increases at the given health care demand $m$. On the other hand, it increases expected utility because the premium decreases (the last two terms in the brackets). Abstracting from a change in $m$ (the last term), expected utility decreases with an increase in the coinsurance rate when the consumer is risk averse.

With respect to waiting time, we find:

$$\frac{\partial EV}{\partial \tau} = \sum_s \pi_s u^c(c^*_s, m^*_s - \theta_s) \left[-f(m^*_s - \theta_s) - (1 - \lambda) \sum_s \pi_s, \frac{\partial m^*_s}{\partial \tau}\right].$$  \hspace{1cm} (15)
To Wait or to Pay for Medical Treatment?

An increase in \( \tau \) decreases income but lowers the premium to the extent that the demand for health services decreases.

**Proposition 1** Full insurance \((\lambda = 0)\) and no insurance coverage \((\lambda = 1)\) are both not optimal.

**Proof**

i) With \( \lambda = 0 \), the optimal allocation is governed by \( u^c = f'(1-\tau)u^m \), and (10) becomes \( \frac{\partial m^*_s}{\partial \lambda} \bigg|_{\lambda=0} = \frac{u^c - u^c m^*_s f'(1-\tau)}{D} < 0 \). Since the consumer chooses the same consumption level across the states of the world, (14) reduces to:

\[
\frac{\partial EV}{\partial \lambda} \bigg|_{\lambda=0} = -u^c \sum_s \pi_s \frac{\partial m^*_s}{\partial \lambda} \bigg|_{\lambda=0} > 0 .
\]

ii) With \( \lambda = 1 \) and \( 0 \leq \tau < 1 \), (14) yields

\[
\frac{\partial EV}{\partial \lambda} \bigg|_{\lambda=1} = \sum_s \pi_s u^c \left[ \sum_s \pi_s m^*_s - m^*_s \right].
\]

The difference in the brackets is positive for small values of \( m^*_s \) and negative for large values of \( m^*_s \). Marginal utility of consumption is negatively correlated to the value of \( m^*_s \) (see (8) and (9)). Let \( m^*_s \) be the maximal value for which the term in the brackets is positive. We then have

\[
u^c \left( c^*_s, m^*_s - \theta_s \right) > u^c \left( c^*_s, m^*_s - \theta_s \right)
\]

for all \( m^*_s > m^*_i \) and

\[
u^c \left( c^*_s, m^*_s - \theta_s \right) < u^c \left( c^*_s, m^*_s - \theta_s \right)
\]

for all \( m^*_s < m^*_i \). Then,

\[
\sum_s \pi_s u^c \left( c^*_s, m^*_s - \theta_s \right) \left[ \sum_s \pi_s m^*_s - m^*_s \right] < u^c \left( c^*_s, m^*_s - \theta_s \right) \sum_s \pi_s \left[ \sum_s \pi_s m^*_s - m^*_s \right].
\]

From \( \sum_s \pi_s = 1 \), the right hand side of this inequality equals zero and thus \( \frac{\partial EV}{\partial \lambda} \bigg|_{\lambda=1} < 0 \).

For an interior solution, \( \lambda \in (0,1) \), we need to check the second derivative. From (14) we obtain:

\[
\frac{\partial^2 EV}{(\partial \lambda)^2} = \sum_s \pi_s \left[ u^c \frac{\partial c^*_s}{\partial \lambda} A_s + u^m \frac{\partial A_s}{\partial \lambda} \right] \quad \text{with} \quad A_s = \sum_s \pi_s m^*_s - m^*_s - (1-\lambda) \sum_s \pi_s \frac{\partial m^*_s}{\partial \lambda}.
\]

---

1. The line of proof follows Breyer et al. (2008).
We find \( \frac{\partial A}{\partial \lambda} = \frac{\partial m^*_s}{\partial \lambda} - (1 - \lambda) \sum_s \pi_s \frac{\partial^2 m^*_s}{(\partial \lambda)^2} \) and from the budget constraint (3) and the first-order condition (6) \( A_s = \frac{\partial c^*_s}{\partial \lambda} + \frac{\partial m^*_s}{\partial \lambda} u^m \). Inserting these expressions into (16) yields:

\[
\frac{\partial^2 EV}{(\partial \lambda)^2} = \sum_s \pi_s u^c \left[ \frac{\partial c^*_s}{\partial \lambda} + \frac{\partial m^*_s}{\partial \lambda} u^c \right] + \sum_s \pi_s u^c \left[ \frac{\partial m^*_s}{\partial \lambda} - (1 - \lambda) \sum_s \pi_s \frac{\partial^2 m^*_s}{(\partial \lambda)^2} \right] .
\]

Given \( u^c, u^m > 0 \), \( u^c < 0 \), \( \partial^2 m^*_s / (\partial \lambda)^2 > 0 \) is a sufficient but not necessary condition for the second-order condition \( (\partial^2 EV / (\partial \lambda)^2 < 0) \) to hold.

Assuming that the second-order condition for an optimal interior solution of the coinsurance rate is fulfilled, we can state:

**Proposition 2** For \( f \geq m^*_s \forall s \), the optimal waiting time is zero.

**Proof** Inserting the condition for an optimal \( \lambda \) (\( \partial EV / \partial \lambda = 0 \)) from (14) into (15) yields:

\[
\frac{\partial EV}{\partial \tau} = -\sum_s \pi_s u^c \left[ f - m^*_s + \sum_s \pi_s m^*_s + (1 - \lambda) \sum_s \pi_s \left[ \frac{\partial m^*_s}{\partial \tau} - \frac{\partial m^*_s}{\partial \lambda} \right] \right] .
\]

For \( f \geq m^*_s \), the difference in the brackets is positive (see (12) for \( u^c > 0 \), \( f^* < 1 \), \( u^c < 0 \), \( \lambda > f^*(1 - \tau) \) and \( D < 0 \)). It follows then that \( f \geq m^*_s \) is sufficient for \( \partial EV / \partial \tau < 0 \).

If \( f \geq m^*_s \), coinsurance not only is more effective in restraining moral hazard than the waiting time. It also involves a lower monetary risk, due to a smaller income effect. If the shock is very high, health care costs may exceed income and the income effect of a marginal increase in the coinsurance rate would be larger than that of a marginal increase in the waiting time. Using coinsurance exclusively may still be optimal. Only if the distribution of \( \theta \) were heavily tailed.
such that excessive risks prevailed would an optimal scheme combine coinsurance with a positive waiting time.\(^2\)

If the coinsurance rate were not optimally chosen, a positive waiting time might also be appropriate. Transforming (15) yields

\[
\frac{\partial EV}{\partial \tau} = -\sum_{r} \pi_{r} u' f \left[ 1 + \frac{(1 - \lambda)}{\tau f} \sum_{r} \pi_{r} \frac{\partial m_{r}}{\partial \tau} \right].
\]

If the elasticity of health demand is comparably high, expected utility varies positively with \(\tau\). An increase in waiting time will then sufficiently reduce health demand so that the decrease in the insurance premium exceeds the income loss.

4. Conclusion

National health services use queuing as a rationing device like centrally planned economies do (Kornai, 1980). Social health insurance systems depend more on user charges to signal the scarcity of resources to the consumers of health care. In this paper, we demonstrate that an optimized design does not generally use a positive waiting time for medical treatment. The reason is that coinsurance sufficiently solves the tradeoff between risk-sharing and moral hazard, thereby rendering waits a redundant instrument. Still, a positive waiting time might be optimal if the choice of the coinsurance rate is restricted or if health status is subject to an excessive risk.

This paper restricts the analysis to a constant coinsurance rate in the tradition of Zeckhauser (1970) and Feldstein and Friedman (1977), among others. With a non-linear scheme, the optimal coinsurance rate tends to be smaller for high health care expenses (Blomquist, 1997). Since for high expenses the optimal waiting time might be positive, it would be interesting to analyze a non-linear coinsurance scheme combined with waiting time, an issue we leave to future research.

\(^2\) If an interior solution for waiting time existed, it could be shown from (14) and (15) that

\[
\frac{\tau^{*}}{\lambda^{*}} = \frac{\sum_{r} \pi_{r} u' \mu_{r}}{\sum_{r} \pi_{r} u' \mu_{m} \sum_{r} m_{r} - \sum_{r} \pi_{r} m_{r}}
\]

with \(\mu_{r}\) and \(\mu_{m}\) as the elasticities of health care demand with respect to waiting time and the coinsurance rate, respectively. This equation corresponds to Eq. (40) in Gravelle and Siciliani (2006), who claimed Ramsey prices and waits for the optimal prioritisation of demand for medical treatment.
Acknowledgments
This paper originated from my discussion of Siciliani and Gravelle (2006) at the 7th European Health Economics Workshop in Constance in May 2006. I thank Peter Zweifel, University of Zurich, and Matthias Kifmann, University of Augsburg, for helpful comments. Suggestions from three anonymous referees have improved the structure of the paper considerably.

5. References
To Wait or to Pay for Medical Treatment?


Zeckhauser, R., 1970. Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives, Journal
Working Paper Series FOR 655

1. Hartmut Kliemt: Priority setting in the age of genomics, December 2007 (1)

2. Marlies Ahlert: If not only numbers count – allocation of equal chances, December 2007 (2)

3. Stefan Felder: The variance of length of stay and the optimal DRG outlier payments, December 2007 (3)


5. Antje Köckeritz: A cooperative bargaining model for two groups of patients, January 2008 (1)

6. Marlies Ahlert and Hartmut Kliemt: Necessary and sufficient conditions to make the numbers count, January 2008 (2)

7. Stefan Felder and Andreas Werblow: Do the age profiles of health care expenditure really steepen over time? New evidence from Swiss Cantons, February 2008 (3)


10. Adele Diederich and Hartmut Kliemt, Public health care priorities at the polls – a note, March 2008 (6)

11. Stefan Felder, To wait or to pay for medical treatment? Restraining ex-post moral hazard in health insurance, April 2008 (7)