WORKING PAPERS

N° 1527

April 2024

"Nonlinear reimbursement rules for preventive and curative medical care"

Helmuth Cremer and Jean-Marie Lozachmeur

Toulouse School of Economics

Nonlinear reimbursement rules for preventive and curative medical care[1]

Helmuth Cremer[2]
and Jean-Marie Lozachmeur[3]

April 2024

[1]Financial support from the Chaire "*Marché des risques et creation de valeur*" of the FdR/SCOR is gratefully acknowledged. We also gratefully acknowledge the funding received by TSE from ANR under grant ANR-17-EURE-0010 (Investissements d'Avenir program).
[2]Toulouse School of Economics, University of Toulouse Capitole, France.
[3]Toulouse School of Economics, CNRS.

Abstract

We study the design of nonlinear reimbursement rules for expenses on secondary preventive and on therapeutic care. With some probability individuals are healthy and do not need any therapeutic health care. Otherwise they become ill and their health status (the severity of their disease) is realized and identifies their *ex post* type. Preventive care is determined *ex ante,* that is before the health status is determined while curative care is chosen *ex post.*

Insurance benefits depend on preventive and curative care in a possibly nonlinear way, and marginal benefits can be positive or negative. In the first best, achieved when health status is *ex post* publicly observable, insurance benefits are flat (lump sum payments) and do not depend on expenditures. When the severity of the disease is not observable, so that there is *ex post* moral hazard, this solution is not incentive compatible (for more healthy individuals). The optimal insurance then implies benefits that increase with *both* types of care. This is because health expenditures reduce informational rents and they are upward distorted. This relaxes the incentive constraint because less healthy individuals value care more than healthy individuals.

Even though preventive care is chosen *ex ante*, when there is no asymmetry of information, it does have an impact on the incentive constraint and thus on informational rents. This is due to two concurring effects. First, prevention is more effective for the more severely ill. Second, these individuals also have a lower marginal utility of income so that a given level of expenditure on preventive care has less impact on their utility.

JEL Codes: I11, I13, I18.

Keywords: *ex post* moral hazard, health insurance, secondary prevention.

1 Introduction

The literature on health insurance design has predominantly concentrated on curative or therapeutic care. This in itself is a complex issue. Even if one abstracts from redistributional considerations and supply side effects associated with imperfect competition, the appropriate insurance coverage is not a trivial problem. Because of asymmetric information there is typically tradeoff between insurance coverage and *ex post* moral hazard which can be mitigated by an appropriate design of reimbursement schemes and the use of copayments. The earlier literature, concentrates on linear reimbursement rules; see for instance Besley (1988). More recent papers consider general, nonlinear policies; see Blomqvist (1997) or Martinon *et al.* (2018).[1]

The insurance coverage of preventive care has received less attention. In practice it varies from country to country and is influenced by the specific healthcare system, the role of public and private insurance, and societal values and priorities.

One can distinguish two types of preventive care. Primary prevention reduces the probability of illness. That is, it aims at reducing occurrence of diseases and health conditions before they develop. Examples include behavioral patterns like exercise, a balanced diet, not smoking or limiting the alcohol intake. Vaccines are another prominent example. While insurers can try to promote primary prevention via counseling and education, insurance coverage *per se* is limited by the fact that it is typically not observable or at least not contractable (verifiable).[2]

Secondary prevention refers to measures aimed at detecting and treating diseases and conditions in their early stages to prevent further progression or complications. In other words, it does not affect the probability of illness but is intended to reduce

[1] For a survey see Zweifel *et al.* (2009), Ch. 6.

[2] Vaccines are at least potentially observable but raise specific problems which go beyond the scope of this paper. In particular, they create a positive externality and even when they are available for free adherence may be too small. And political considerations often imply that mandates are not a realistic option.

its severity. Examples include checkups and diagnostic screening like mammographies colonoscopies, pap smears, blood tests, and other tests that can detect cancer, diabetes, heart disease, and other health conditions. Secondary prevention is typically observable and verifiable so that it can be covered by an insurance scheme.

We study the design of reimbursement rules of preventive and curative (therapeutic) care. Because most of primary prevention is not verifiable we concentrate on secondary prevention.[3] The main contribution of this paper is that it considers nonlinear policies. In other words, we determine the best policy given the information available to the insurer. Most of the existing literature restricts policies to be linear (affine). This is an *ad hoc* assumption which is not justified by the information structure. For curative care, a notable exception is Blomqvist (1997) who studies nonlinear reimbursement rules, albeit for a somewhat restrictive utility function. To our knowledge the few papers who consider preventive care all restrict policies to be linear. The most noticeable example is Barigozzi (2004) who also considers secondary prevention along with therapeutic care. As will become clear below the generalization to nonlinear rules is not just of methodological interest; it also has a drastic impact on the results. With linear rules, Barigozzi shows that while treatment expenses should always be subsidized this is true for prevention if and only if prevention reduces the cost of treatment, that is in the case the two activities are substitutes. We will show that this result is an artifact of the linearity assumption. With nonlinear schemes both types of care they should be subsided (at the margin) irrespective of the substitutability or complementarity of prevention and treatment. Intuitively, linear copayments mechanically create substitution effects (direct and across types of care) which can be avoided with nonlinear policies. When reimbursement rules are restricted solely by the available information, they have to be designed according to their impact on informational rents.

Our model considers is a large number (or a continuum) of *ex ante* identical individ-

[3] Ellis and Manning (2007) also consider prevention and treatment but concentrate on primary prevention.

uals who are endowed with a given disposable income. With some probability they are healthy and do not need any curative (therapeutic) health care. Otherwise they become ill and their health status (the severity of their disease) is realized and identifies their *ex post* type. Preventive care is determined *ex ante,* that is before the health status is determined while therapeutic care is chosen *ex post.*

We study the design of a social insurance scheme that maximizes individual's expected utility subject to the resource constraint, which requires that total contributions (payroll taxes or premium payments) equal expected health insurance benefits. Insurance benefits depend on preventive and curative care in a possibly nonlinear way, and marginal benefits can be positive or negative. We first study the case where the health status is *ex post* publicly observable which yields the first best optimum. Then we turn to the case where severity of disease is not observable to the insurer. Throughout the paper we assume that expenditures on preventive care and health status (healthy or sick) are observable at the individual level.

In the first best insurance benefits are flat (lump sum payments) and do not depend on expenditures. When the severity of the disease is not observable, this solution cannot be implemented because individuals in good health would mimick the less healthy individuals. The optimal insurance implies benefits that increase with both types of care. This is because health expenditures reduce informational rents and they are upward distorted. For therapeutic care this generalizes the result of Blomqvist (1997) and the intuition is easily understood. Less healthy individuals value care more than healthy individuals. Consequently, an increase in expenditures on (therapeutic) care relaxes the incentive constraint.

The case of preventive care is more complex. One might at first be tempted to think that a solution would leave the choice of preventive care undistorted and just provide a flat payment like under full information. Indeed, preventive care is chosen *ex ante*, at a point where there is uncertainty but no asymmetric information. Individuals and insurers alike do not know the (future) realization of the state of health and the severity in case of disease. Consequently it is not immediately obvious what positive effect a distortion might bring about. Our formal analysis shows that this first intuition is misleading; prevention does have an impact on the incentive constraint and thus on informational rents. Specifically as preventive care increases, utility decreases less fast with the severity of the disease so that the rents enjoyed by healthier individuals decrease. Intuitively this is due to two concurring effects. First, prevention is more effective for the more severely ill. Second, these individuals also have a lower marginal utility of income so that a given level of expenditure on preventive care has less impact on their utility.

2 The model

2.1 Individuals

The is a large number (or a continuum) of *ex ante* identical individuals who are endowed with a disposable income ω. With probability π, they are healthy (state of nature H) and do not need any curative (therapeutic) health care. The utility of healthy individuals is given by $v(c_0)$, where c_0 is net consumption; assume $v'(c_0) > 0$ and $v''(c_0) < 0$.

With probability $(1-\pi)$ they become ill (state of nature S) and their health status (the severity of their disease) is represented by a parameter θ, which is also used to identify their *ex post* type. The random variable θ is distributed over $\Theta \equiv [\underline{\theta}, \overline{\theta}] \subset \Re_+$with a density $f(\theta)$ and a distribution function $F(\theta)$. Note that a larger value of θ corresponds to a more severe disease thus a worse health status. Individual of type θ has preferences

$$u(c, m, e, \theta) \tag{1}$$

where c denotes consumption of a numeraire good, m medical expenditures (curative care) and e secondary prevention expenditures. We will be more specific on the timing below, but it important to note from the outset that preventive care is determined *ex* *ante,* that is before the health status is determined while m is chosen *ex post.* Consequently, m can be conditioned on θ while e is by definition the same in all states of nature. An individual's expected utility is thus given by

$$EU = \pi v\left(c_0\right) + (1-\pi)\int_{\underline{\theta}}^{\overline{\theta}} u[c\left(\theta\right), m\left(\theta\right), e, \theta]f\left(\theta\right)d\theta \tag{2}$$

We assume $u_c > 0$, $u_m > 0$, $u_e > 0$ so that consumption as well as both types of medical care increase utility for any level of θ, and $u_{cc} < 0$ which implies that individuals are risk averse.[4] Furthermore, we have $u_\theta < 0$ which reflects the assumption that a larger θ corresponds to a more severe disease. Furthermore, we assume $u_{m\theta} > 0$ so that the benefits of medical care increase with the severity of the illness. Consequently, absent of any insurance, individuals with a larger θ choose a larger level of m.[5] Finally, we suppose that $u_{c\theta} \leq 0$. In words, the marginal utility of net income decreases with the severity of the illness. Empirically, a strict decrease appears to be the most plausible assumption; see Finkelstein et al. (2013). However, some of the literature, including Blomqvist (1997), assumes $u_{c\theta} = 0$ for tractability. Consequently, we did not want to rule out this special case.[6]

2.2 Policy design

We study the design of a social insurance scheme that maximizes individual's expected utility subject to the resource constraint. This constraint requires that total contributions (payroll taxes or premium payments) equal expected health insurance benefits.

[4] Subscripts refer to partial derivatives.

[5] Absent of any insurance we have

$$\frac{\partial m}{\partial \theta} = -\frac{u_{m\theta}}{SOC},$$

where $SOC < 0$ is the second-order condition for an interior solution which is assumed to hold.

[6] Barigozzi (2004) uses a utility function given by $u\left(c + H\left(e, m\right)\right)$; there are just two states of nature corresponding to our S and H and the severity of illness is not considered. The preferences considered by Besley (1988), Cremer and Lozachmeur (2022) and by Martinon et al, (2018) are also encompassed by (1); they account for the severity θ but do not have prevention.

Social insurance "covers" both preventive and curative care. To be more precise, benefits depend on e and m in a possibly nonlinear way and marginal benefits can be positive or negative. We first study the case where the health status θ is *ex post* publicly observable. Then we turn to the case where individuals θ's are not observable to the insurer. Throughout the paper we assume that e and the health status, H or S are observable at the individual level.

Note that while we focus on social insurance the same equilibrium would emerge in a private insurance market with identical insurers, perfect competition and free entry. In equilibrium, profits are zero; there is no loading factor. Under these assumptions the problem of a private insurer is to maximize the expected utility of the representative individual under a zero profit constraint, which is exactly the same as that of welfare maximizing social insurance.

2.3 Timing

Formally a policy consists of a premium P and a benefit function

$$B(e, m(\theta)) = I(e) + R\left(m\left(\theta\right)\right), \tag{3}$$

where $I(e)$ is the (positive or negative) reimbursement of preventive care while $R\left(m\left(\theta\right)\right)$ is the transfer associated with curative care expenditures. Note that splitting B into two part I and R is done only for the ease of exposition and has no impact on the results.

The timing is as follows. First, the social insurer announces the policy $\{P, I\left(e\right), R(m(\theta)\}$, specifying the premium P, paid *ex ante* and the benefit rule defined in (3). Second, individuals choose their level of preventive care e. Note that because preventive care is by definition determined *ex ante* the same level is chosen by all individuals and it cannot be conditioned on θ. Third, the state of nature is realized and the variable θ is drawn for all individual in state S and revealed to them. Finally, individuals in state S choose their level of health care expenses $m\left(\theta\right)$ depending on their health status θ.

2.4 Individual problem

To determine the optimal reimbursement policy we shall use a mechanism design approach and determine first the allocation that is induced by this policy. To examine how the optimal policy can be implemented by the considered instruments we have to study an individual's problem. For a given policy, that is a premium P and a reimbursement policy $I(e)$ and $R(m(\theta))$ agents choose e and $m(\theta)$ by solving the following problem

$$\max_{e,m(\theta)} \pi v[\omega - P + I(e) - e] + (1-\pi) \int_{\underline{\theta}}^{\overline{\theta}} u[\omega - P - m(\theta) - e + I(e) + R(m(\theta)), m(\theta), e, \theta] f(\theta) \, d\theta.$$

Differentiating with respect to e and $m(\theta)$ and rearranging yields

$$MRS_{cm} = \frac{u_m}{u_{c(\theta)}} = 1 - R'(m(\theta)) \qquad \forall \theta \in \Theta, \tag{4}$$

$$MRS_{ce} = \frac{(1-\pi) \int_{\underline{\theta}}^{\overline{\theta}} u_e f(\theta) \, d\theta}{\pi v'(c_0) + (1-\pi) \int_{\underline{\theta}}^{\overline{\theta}} u_{c(\theta)} f(\theta) \, d\theta} = 1 - I'(e). \tag{5}$$

Condition (4) states that the marginal benefit of m, expressed in monetary terms must equal its marginal cost accounting for the reimbursement. Since θ is chosen *ex post* no uncertainty is involved. The interpretation of (5) concerning e is similar. Note that since e is chosen *ex ante* its benefits are uncertain and depend on the realization of H and θ. When $R(m(\theta)) = I(e) = 0$ and $R'(m(\theta)) = I'(e) = 0$ for all levels of m and e we obtain the *laissez-faire* solution with no insurance.

For future reference also note that

$$\frac{\partial MRS_{cm}}{\partial \theta} = \frac{u_c u_{m\theta} - u_{c\theta} u_m}{(u_c)^2} > 0. \tag{6}$$

In words at any given point in the (m, c) space, individuals with a larger θ (who are in worse health) have steeper indifference curves and thus a higher willingness to pay for

m. This, in turn implies the single crossing property of indifference curves in the (m, c) plane.

3 The full information optimum

Our main focus is of course on the policy design when θ is not observable. To understand its properties, the full information optimum provides and interesting benchmark.

Define $d_0 = c_0 + e = \omega + I - P$ and

$$d(\theta) = c(\theta) + m(\theta) + e = \omega + R(\theta) + I - P. \tag{7}$$

Intuitively, $d(\theta)$ denotes the total resources available to an individual in state S and of type θ, including reimbursement of medial care net of the premium. This budget is allocated to consumption and both types of medical care. The variable d_0 has a similar interpretation for an individual in state H for whom it is allocated to consumption and preventive care. For consistency with the solution under asymmetric information we use d_0, $d(\theta)$, $m(\theta)$ and e as decision variables. The problem of the social planner is

$$\max_{d_0, d(\theta), m(\theta), e} \pi v(d_0 - e) + (1-\pi) \int_{\underline{\theta}}^{\overline{\theta}} u[d(\theta) - m(\theta) - e, m(\theta), e, \theta] f(\theta) d\theta \tag{8}$$

$$\text{s.t.} \quad \omega - \pi d_0 - (1-\pi) \int_{\underline{\theta}}^{\overline{\theta}} d(\theta) f(\theta) d\theta \geq 0, \tag{9}$$

In words, we maximize expected utility of a representative individual subject to the resource constraint. Not that the reimbursement policy does not *explicitly* appear in this problem but is implicitly defined by (7) together with (9) and the definition of d_0. Note that combining these two equations we obtain

$$\pi(I-P) + (1-\pi) \int_{\underline{\theta}}^{\overline{\theta}} (R(\theta) + I - P) f(\theta) = I - P + (1-\pi) \int_{\underline{\theta}}^{\overline{\theta}} R(\theta) f(\theta) = 0,$$

so that the budget of the insurer is balanced.

Denoting by μ the multiplier associated with the government budget constraint and differentiating the Lagrangrian expression $\mathcal{L}$ yields the following first-order conditions (FOCs).

$$\frac{\partial \mathcal{L}}{\partial d_0} = \pi[v'(c_0) - \mu] = 0, \tag{10}$$

$$\frac{\partial \mathcal{L}}{\partial d(\theta)} = (1-\pi)[u_{c(\theta)} - \mu]f(\theta) = 0, \quad \forall \theta \in \Theta \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial m(\theta)} = (1-\pi)[u_m - u_{c(\theta)}]f(\theta) = 0, \quad \forall \theta \in \Theta \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial e} = (1-\pi)\int_{\underline{\theta}}^{\overline{\theta}} u_e f(\theta)\, d\theta - [\pi v'(c_0) + (1-\pi)\int_{\underline{\theta}}^{\overline{\theta}} u_{c(\theta)} f(\theta)\, d\theta] = 0. \tag{13}$$

Combining (10) and (11) yields

$$v'(c_0) = u_{c(\theta)} \quad \forall \theta \in \Theta,$$

so that marginal utilities of income are equalized in all states of nature. In other words, individuals are fully insured. Furthermore, (12) and (7) imply

$$MRS_{cm} = \frac{u_m}{u_{c(\theta)}} = 1, \tag{14}$$

which from (4) requires

$$R'(m(\theta)) = 0 \quad \forall \theta \in \Theta. \tag{15}$$

Similarly (13) and (7) imply

$$MRS_{ce} = \frac{(1-\pi)\int_{\underline{\theta}}^{\overline{\theta}} u_e f(\theta)\, d\theta}{\pi v'(c_0) + (1-\pi)\int_{\underline{\theta}}^{\overline{\theta}} u_{c(\theta)} f(\theta)\, d\theta} = 1 \tag{16}$$

so that from (5) we have

$$I^{'}(e) = 0. \tag{17}$$

In words, (15) and (17) mean that the *marginal* reimbursement of expenditures on medical care e and $m(\theta)$ are equal to zero. Consequently medical care levels are not distorted: their marginal benefits, as measured by the marginal rates of substitution in (14) and (16), are equal to their marginal costs (namely 1).

These properties are not surprising. With θ observable and absent any *ad hoc* restrictions on instrument the solution is of course first-best efficient. This, in turn requires full insurance and undistorted medical expenses.

Note that with purely linear instruments this would of course not be possible. But with nonlinear instruments and given full information we can give each individual the appropriate *flat* benefit, which does not directly depend on the individual's expenditures on medical care.

The main results of this section are summarized in the following proposition.

Proposition 1 *When there is full information so that the health status and the severity of the illness are observable and nonlinear instruments are available the optimal solution implies*

(i) Full insurance, so that marginal utility of income is equalized across states of nature.

(ii) The reimbursement rules of curative and preventive care are flat; marginal reimbursement rates are zero: $R'(m) = I'(e) = 0$.

(iii) Medical care levels are not distorted: their marginal benefits are equal to their marginal costs.

4 Solution under asymmetric information

Providing each individual with the appropriate flat benefit is of course only possible when θ is publicly observable. When this is not the case the reimbursement policy has to be based on observable variables and specifically $m(\theta)$ and e. Since e is chosen *ex ante*, it is not immediately obvious that the information asymmetry would be relevant. By contrast, as far as m is concerned, this is obvious. When the reimbursement is based on the individual's level of expenditure there is an obvious problem of *ex post* moral hazard an issue which is well known in the health economics literature. Specifically a positive reimbursement rate will tend lead to excessive consumption of care.

When θ is not observable, the insurance policy must be incentive compatible that is *ex post* all individuals must prefer their own consumption bundle to that available to any other type. As usual we solve this problem by first deriving the best incentive compatible allocation and then study how it can be implemented by an insurance policy specifying the reimbursement rules of medical care.

4.1 The problem

We continue to use the same decision variables as in the previous section namely d_0, $d(\theta)$, $m(\theta)$ and e. The problem of the social planner is now given by

$$\max_{d_0,d(\theta),m(\theta),e} \pi v\left(d_0 - e\right) + (1-\pi)\int_{\underline{\theta}}^{\overline{\theta}} u[d\left(\theta\right) - m\left(\theta\right) - e, m\left(\theta\right), e, \theta] f(\theta) d\theta \tag{18}$$

$$\text{s.t.} \quad \omega - \pi d_0 - (1-\pi)\int_{\underline{\theta}}^{\overline{\theta}} d\left(\theta\right) f\left(\theta\right) d\theta \geq 0, \tag{19}$$

$$\begin{aligned} &u\left(d\left(\theta\right) - m\left(\theta\right) - e, m\left(\theta\right), e, \theta\right) \geq \\ &u\left(d\left(\theta'\right) - m\left(\theta'\right) - e, m\left(\theta'\right), e, \theta\right) \qquad \forall\ \theta, \theta' \in \Theta \end{aligned} \tag{20}$$

which differs from the problem under full information (8)–(9) in that we have added an incentive constraint for each type θ, equation (20).

4.2 The local incentive constraint

To solve this problem we use a first-order approach which leads us to consider a relaxed problem. Specifically we consider a direct mechanism consisting for a bundle d_0 $\{d(\theta), m(\theta)\}$ for each θ. Individuals choose their reported type θ' which maximizes their utility given their true type and the policy $\{d(\theta), m(\theta)\}$. Formally, they solve

$$\max_{\theta'} \quad \pi v(d_0 - e) + (1-\pi) u\left(d\left(\theta'\right) - m\left(\theta'\right) - e, m\left(\theta'\right), e, \theta\right). \tag{21}$$

Using the FOC associated with this problem the local incentive constraint can be written as

$$\dot{U}(\theta) = (1-\pi) u_\theta (d(\theta) - m(\theta) - e, m(\theta), e, \theta), \tag{22}$$

where $U(\theta) = \pi v(d_0 - e) + (1-\pi)[d(\theta) - m(\theta) - e, m(\theta), e, \theta]$ and where $\dot{U}(\theta)$ denotes the total derivative of U with respect to θ.[7]

The local approach is valid if the second-order condition of problem (21) is satisfied for which a sufficient condition is $\dot{m}(\theta) > 0$. For simplicity we assume that this property holds in equilibrium.[8]

Observe that equation (22) implies $\dot{U}(\theta) < 0$ so that utility *decreases* with θ. In other words, it *increases* as θ *decreases*. Note that the faster U decreases with θ, the larger are the rents enjoyed by more healthy individuals.

[7] The FOC is given by

$$[\dot{d}(\theta') - \dot{m}(\theta')] u_c [d(\theta') - m(\theta') - e, m(\theta'), e, \theta] + \dot{m}(\theta') u_m \left(d\left(\theta'\right) - m\left(\theta'\right) - e, m\left(\theta'\right), e, \theta\right) = 0, \tag{23}$$

and to derive (22) we have used the fact that to ensure truthful revelation it must be satisfied for $\theta' = \theta$. Intuitively this amounts to using the enveloppe theorem whic himplies that the total derivative of U with repect to θ is equal to the partial derivative.

[8] If it violated than the solution involves bunching over some interval(s). As in much of the literature on contract theory and in particular optimal taxation, we neglect this complication as it adds little to the understanding of the underlying economic intuition.

4.3 The relaxed problem

As usual in contract theory we determine the solution by using an optimal control approach with $U(\theta)$ as state variable and $m(\theta)$ as control, while also optimizing with respect to e which being set *ex ante* is not contingent on θ. To be consistent with standard optimal control theory we add the control variable $z(\theta)$ and impose the constraint that $z(\theta) = e$.

Formally, the problem of the government can then be stated as follows

$$\max_{U(\theta),m(\theta),z(\theta),e} \int_{\underline{\theta}}^{\overline{\theta}} U(\theta)\, f(\theta)\, d\theta \tag{24}$$

$$\text{s.t.} \quad \omega - \pi d_0 - (1-\pi) \int_{\underline{\theta}}^{\overline{\theta}} d(\theta)\, f(\theta)\, d\theta \geq 0 \tag{25}$$

$$U(\theta) = \pi v(d_0 - e) + (1-\pi)\, u(d(\theta) - m(\theta) - z(\theta), m(\theta), e, \theta) \tag{26}$$

$$z(\theta) - e = 0 \tag{27}$$

$$\dot{U}(\theta) = (1-\pi)\, u_\theta (d(\theta) - m(\theta) - e, m(\theta), e, \theta) \tag{28}$$

The Hamiltonian associated with this problem is

$$\begin{aligned}
\mathcal{H} &= f(\theta)\, U(\theta) \\
&\quad - \alpha(\theta) \left[U(\theta) - \pi v(d_0 - e) - (1-\pi)\, u(d(\theta) - m(\theta) - z(\theta), m(\theta), z(\theta), \theta)\right] \\
&\quad + \mu \left[\omega - \pi d_0 - (1-\pi)\, d(\theta)\right] f(\theta)\, d\theta \\
&\quad - \eta(\theta)[z(\theta) - e] \\
&\quad + \lambda(\theta)\, (1-\pi)\, u_\theta (d(\theta) - m(\theta) - z(\theta), m(\theta), z(\theta), \theta)
\end{aligned} \tag{29}$$

Where $\lambda(\theta)$ is the costate variable associated with equation (28) while μ, $\alpha(\theta)$ and $\eta(\theta)$ are the Lagrange multipliers associated respectively with constraints (25), (26) and (27). Differentiating $\mathcal{H}$ with respect to $d_0, d(\theta)$, $m(\theta)$, $z(\theta)$, e and applying Pontryagin's maximum principle yields the following necessary conditions defining the solution[9]

$$\alpha(\theta)\pi v'(d_0 - e) - \mu\pi f(\theta) = 0, \tag{30}$$

$$\alpha(\theta)(1-\pi)u_c - (1-\pi)\mu f(\theta) + \lambda(\theta)(1-\pi)u_{c\theta} = 0, \tag{31}$$

$$\alpha(\theta)(1-\pi)(u_m - u_c) + \lambda(\theta)(1-\pi)(u_{\theta m} - u_{c\theta}) = 0, \tag{32}$$

$$-\eta(\theta) + \alpha(\theta)(1-\pi)(u_e - u_c) + \lambda(\theta)(1-\pi)(u_{\theta e} - u_{\theta c}) = 0, \tag{33}$$

$$-\alpha(\theta)\pi v'(d_0 - e) + \eta(\theta) = 0, \tag{34}$$

$$\dot{\lambda}(\theta) = -\frac{\partial \mathcal{H}}{\partial U(\theta, e)} = \alpha(\theta) - f(\theta). \tag{35}$$

and the transversality conditions are given by:

$$\lambda(\underline{\theta}) = \lambda(\overline{\theta}) = 0. \tag{36}$$

Conditions (34)–(36) along with the resource constraint (25) define the optimal allocation. From (31), one has

$$\alpha(\theta) = \frac{\mu f(\theta)}{u_c} - \lambda(\theta)\frac{u_{c\theta}}{u_c} \tag{37}$$

Substituting in (32) and rearranging successively yields

$$\left(\frac{\mu f(\theta)}{u_c} - \lambda(\theta)\frac{u_{c\theta}}{u_c}\right)(u_m - u_c) + \lambda(\theta)(u_{\theta m} - u_{c\theta}) = 0,$$

$$\mu f(\theta)\frac{u_m}{u_c} - \mu f(\theta) - \lambda(\theta)u_{c\theta}\frac{u_m}{u_c} + \lambda(\theta)u_{c\theta} + \lambda(\theta)(u_{\theta m} - u_{c\theta}) = 0,$$

$$\mu f(\theta)\left(\frac{u_m}{u_c} - 1\right) + \lambda(\theta)\left(u_{\theta m} - u_{c\theta}\frac{u_m}{u_c}\right) = 0. \tag{38}$$

Combining equation (38) with the individual's first-order condition under the implementing benefit rule (4), one obtains

$$R'(m) = \frac{\lambda(\theta)\left(u_{\theta m} - u_{c\theta}\frac{u_m}{u_c}\right)}{\mu f(\theta)}. \tag{39}$$

[9] See for instance Takayama (1985), pages 602–603.

Recall that $u_{\theta m} > 0$ and $u_{\theta c} \geq 0$ and based on the usual properties of Lagrangian multipliers we have $\mu > 0$. In Appendix A we also show that $\lambda(\theta) > 0$ for $\theta \in]\underline{\theta}, \overline{\theta}[$. Consequently, equation (39) implies that $R'(m(\theta)) > 0$ for $\theta \in]\underline{\theta}, \overline{\theta}[$. Together with $\dot{m}(\theta) > 0$ this means that $R'(m) > 0$ for $m \in]m(\underline{\theta}), m(\overline{\theta})[$. In words the marginal reimbursement rate of curative health care is positive except at the endpoints of the interval. This implies that compared to the efficient, full information, outcome health care is distorted upwards. Intuitively, this distortion is explained by the usual rent reduction effect. To understand this recall that in this setting mimicking goes form low θ's (more healthy individuals) to higher θ's (less healthy individuals). Furthermore, equation (6) implies that individuals with a higher θ (the mimicked) have a larger willingness to pay than the mimicking individual with a lower θ. Consequently the upward distortion relaxes the otherwise binding incentive constraint so that the rents of healthy individuals are reduced.

This result is rather intuitive and in line with standard properties obtained in optimal tax models.[10] A similar result was already obtained by Blomqvist (1997) and our contribution regarding m is mainly that we generalize Blomqvist's analysis. For practical purposes this means that $R'(m) > 0$ is a very robust result and does not rely on the specific assumptions imposed by Blomqvist (like the separability).

Anyway, the main focus of our paper is preventive care to which we now turn. One might at first be tempted to think that a solution would leave the choice of e undistorted and just provide a flat payment like under full information. Indeed, preventive care is chosen *ex ante*, at a point where there is uncertainty but no asymmetric information. Individuals and insurers alike do not know the (future) realization of the state of health and in state S, the severity θ. Consequently it is not immediately obvious what positive effect a distortion might bring about.

[10] Except that all signs are reversed because a high θ refers to the "bad" type so that the incentive constraint binds upwards. In optimal tax models by contrast a large w corresponds to the "good" type and the downward incentive constraint is binding.

Our formal analysis show that this first intuition is misleading; prevention does have an impact on the incentive constraint and thus on informational rents. We first establish this result formally and then further discuss the intuition.

Substituting (34) in (33) yields:

$$-\alpha(\theta)\pi v'(d_0 - e) + \alpha(\theta)\left(1-\pi\right)\left(u_e - u_c\right) + \lambda\left(\theta\right)\left(1-\pi\right)\left(u_{\theta e} - u_{\theta c}\right) = 0$$

Dividing by $\alpha\left(\theta\right)$ and multiplying by $f\left(\theta\right)$ yields :

$$-\pi v'(d_0 - e) f\left(\theta\right) + \left(1-\pi\right)\left(u_e - u_c\right) f\left(\theta\right) + \frac{\lambda\left(\theta\right)}{\alpha\left(\theta\right)}\left(1-\pi\right)\left(u_{\theta e} - u_{\theta c}\right) f\left(\theta\right) = 0$$

Recall that $u_{\theta c} < 0$ while $u_{\theta e} > 0$ so that $\left(u_{\theta e} - u_{\theta c}\right) > 0$. Integrating over θ and rearranging yields

$$\pi v'(d_0 - e) + \left(1-\pi\right)\int_{\underline{\theta}}^{\overline{\theta}} u_c f\left(\theta\right) d\theta = \left(1-\pi\right)\int_{\underline{\theta}}^{\overline{\theta}} u_e f(\theta) d\theta + \int_{\underline{\theta}}^{\overline{\theta}} \frac{\lambda\left(\theta\right)}{\alpha\left(\theta\right)}\left(1-\pi\right)\left(u_{\theta e} - u_{\theta c}\right) f\left(\theta\right) d\theta.$$

Dividing by $EU_c = \left[\pi v'\left(d_0 - e\right) + \left(1-\pi\right)\int_{\underline{\theta}}^{\overline{\theta}} u_c f\left(\theta\right) dx\right]$ and using (5) implies:

$$I'\left(e\right) = \frac{\left(1-\pi\right)\int_{\underline{\theta}}^{\overline{\theta}} \frac{\lambda(\theta)}{\alpha(\theta)}\left(u_{\theta e} - u_{\theta c}\right) f\left(\theta\right) d\theta}{EU_c},$$

which using (34) to substitute for $f(\theta)/\alpha(\theta)$ implies

$$I'\left(e\right) = \frac{\left(1-\pi\right) v'\left(c_0\right)}{\mu} \frac{\int_{\underline{\theta}}^{\overline{\theta}} \lambda\left(\theta\right)\left(u_{\theta e} - u_{\theta c}\right) d\theta}{EU_c} > 0. \tag{40}$$

Consequently, preventive care must be subsidized at the margin which in turn implies an upward distortion compared to the full information solution. The numerator on the right-hand-side of this expression measures the welfare benefit of an increase in e via its impact on the incentive constraint. To understand this note that $\left(u_{\theta e} - u_{\theta c}\right) > 0$ is the derivative of (28) with respect to e (which affects the first and the third argument of u_θ). Since it is positive it means that $\dot{U}$ increases, which since $\dot{U} < 0$ implies that the absolute value of $\dot{U}$ decreases. Consequently, utility decreases less fast as θ increases so that the rents enjoyed by healthier individuals decrease. To sum up, even though e is chosen *ex ante*, it affects the rents enjoyed by healthier individuals *ex post*.[11] Intuitively this is due to two concurring effects. First, prevention is more effective for the more severely ill and second, these individuals also have a lower marginal utility of income so that a given level of e has less impact on their utility.

A striking feature of this result is that it does *not* depend on the sign of u_{me} which can be interpreted as the degree of complementarity between preventive and curative care. More precisely, when $u_{me} > 0$ the two types of care can be consider as complements (prevention makes treatment more effective) while they are substitutes when $u_{me} < 0$ (so that the marginal benefit of curative care decreases with the level of prevention). This is in sharp contrast to the results obtained in linear models and particularly to Barigozzi (2004) who shows that the sign of I' crucially depends on the degree of complementarity. In her setting it is possible that preventive care is taxed.[12] Our analysis shows that these results are merely an artifact of the linearity assumption which in turn is *ad hoc* and in no way justified by informational considerations. When instruments are restricted solely by the information structure and can be nonlinear the results are clear and simple in the sense that both types of care should be subsidized at the margin.

Finally let us return to the extent of insurance coverage. We have shown that, as expected, the full information solution involves full insurance in the sense that the marginal utility of income is equalized across states of nature. This implies that both the risk of illness and of its severity are fully insured. In that context the result was easily obtained and followed directly from equations (10) and (11). Under asymmetric information, the counterparts to these conditions are (30) and (31). Because (31) depends on the incentive constraint (via the third term) a simple inspection of the expressions

[11] The denominator of (40) is simply the expected marginal utility of consumption which normalizes the welfare impact to express it in monetary terms. This makes it comparable to I' which is also in monetary terms.

[12] More precisely both types of care can be taxed or subsidized.

shows that they no longer (directly) imply $v' = u_{c(\theta)}$ for all $\theta \in \Theta$ and there is no reason to believe that this property would hold in general. And indeed equation (37) together with (30) implies

$$\left[v'(c_0) - u_{c(\theta)}\right] = \lambda(\theta)\, u_{c\theta}/\alpha(\theta) < 0 \text{ for all } \theta$$

so that $v'(c_0) < u_{c(\theta)}$. Consequently, there is underinsurance for the risk of being ill that is for the state S.

Turning to the severity of illness, the extent of insurance coverage is less obvious. Indeed, it is not clear whether u_c is increasing or decreasing in θ. Observe that an increasing u_c would reflect underinsurance while a decreasing profile would involve overinsurance. The first-order condition (23) in footnote 7 implies that $c(\theta)$ is decreasing which everything else equal would imply that u_c is increasing.[13] However, there are other effects and in particular the one associated with assumption that $u_{c\theta} < 0$, which tends to make u_c decreasing. This effect disappears if as Blomqvist (1997) one assumes that $u_{c\theta} = 0$, but as discussed above (and acknowledged by the author) this is not a realistic assumption.[14] However, it follows by continuity that when the absolute value of $u_{c\theta}$ is sufficiently small, there is also underinsurance for the severity of illness. By contrast when this cross derivative is large (in absolute value) the possibility of overinsurance cannot be ruled out. To sum up, while it is clear that the solution does not in general involve full insurance or the severity of disease, it does not appear to be possible to determined whether it involves over- or underinsurance without making further assumptions on the utility function.

The main results of this section are summarized in the following proposition.

[13] To see this observe that simplifying notation, equation (23) can be written as

$$(\dot{d} - \dot{m})u_c + \dot{m}u_m = 0,$$

so that $(\dot{d} - \dot{m}) < 0$, which implies that $c(\theta) = d(\theta) - m(\theta) - e$ is decreasing.

[14] And even with this assumption one cannot make a definitive conclusion because m increases with θ which in turn might affect u_c, and we haven't made any assumption regarding u_{cm}.

Proposition 2 *When the severity of the illness is not publicly observable and nonlinear instruments are available the optimal solution*

(i) Does not imply full insurance; marginal utility of income is not in general equalized across states of nature. The risk of illness (being in state S) is underinsured. Regarding the severity of the disease (the realization of θ), when $u_{c\theta}$ is zero or sufficiently small (in absolute value), there is underinsurance. But when (the absolute value of) $u_{c\theta}$ is large the result is ambiguous and no general conclusion can be reached without further restrictions on the utility function.

(ii) Implies a marginal subsidy on both types of care so that $R'(m(\theta)) > 0$ (except at the endpoints of the support of θ when $R'(m) = 0$), and $I'(e) > 0$. In other words m and e are distorted upwards, irrespective of the degree of complementarity between preventive and curative care.

(iii) In both cases the distortion is imposed to mitigate rents (relax the incentive constraint).

(iii)a For m this is intuitively explained by the relative slopes of the mimicked and the mimicker's indifference curves in the (m, c) space, exactly like in an optimal income tax model.

(iii)b Since e is chosen ex ante and the same for all the effects at work are more complex. A larger e provides benefits that increase with the severity of the illness. Consequently utility decreases less fast as θ increases so that the rents enjoyed by healthier individuals decrease.

5 Conclusion

We have studied the design of nonlinear reimbursement rules of preventive and curative (therapeutic) care. We have concentrated on secondary prevention which is typically verifiable. Most of the existing literature restricts policies to be linear (affine). By contrast, we determine the best policy given the information available to the insurer without imposing such an *ad hoc* assumption. This has a drastic impact on the results. With linear rules, prevention should be subsidized if and only it reduces the cost of treatment, that is when the two types of care are substitutes. With nonlinear schemes *both* types of care there should be subsided (at the margin) irrespective of the substitutability or complementarity of prevention and treatment.

We have shown that in the first best (when the severity of illness is observable) insurance benefits are flat (lump sum payments) and do not depend on expenditures. When the severity of the disease is not observable, there is *ex post* moral hazard and this solution is not incentive compatible. The optimal insurance implies benefits that increase with *both* types of care. This is because health expenditures reduce informational rents and they are upward distorted. For therapeutic care this generalizes the result of Blomqvist (1997) and the intuition is easily understood. Less healthy individuals value care more than healthy individuals. Consequently, an increase in expenditures on (therapeutic) care relaxes the incentive constraint.

The case of preventive care is more complex because preventive care is chosen *ex ante*, at a point where there is uncertainty but no asymmetric information. We have shown that prevention nevertheless does have an impact on the incentive constraint and thus on informational rents. Specifically as preventive care increases, utility decreases less fast with the severity of the disease so that the rents enjoyed by healthier individuals decrease. Intuitively this is due to two concurring effects. First, prevention is more effective for the more severely ill. Second, these individuals also have a lower marginal utility of income so that a given level of expenditure on preventive care has less impact on their utility.

Finally, while the first best implies full insurance the second best does not. The risk of disease is underinsured while no general conclusion regarding insurance coverage can be reached without further restrictions on utility. In particular when the marginal utility of decreases sufficiently fast with the severity, overinsurance cannot be ruled out.

We have ignored a number of potentially relevant issues that might affect insurance design. In particular, we have not considered *ex ante* income heterogeneity.[15] Clearly, the insurance coverage of health care involves many redistributive issues. In particular, subsidizing preventive care can also help promote health equity by making these services more accessible to a broader range of individuals, regardless of their income or financial situation. The redistributive role of health insurance (to supplement taxation) has been studied by Rochet (1991) and Cremer and Pestieau (1996). These papers have shown the complexity of the underlying problem because it involves multidimensional heterogeneity. Either way neither of these papers considers preventive care.

Insurance may also take into account age and risk factors when determining coverage for secondary prevention. For example, certain screenings may be recommended at specific ages or for individuals with known risk factors, and insurance should cover these as appropriate. While conditioning coverage of curative care on observable risk factors is equity perspective and indeed typically ruled out by anti-discrimination laws, encouraging screening test for specific risk groups is common practice. Formally this would amount to introducing tagging into our setting.

Finally, we have assumed that individuals are not myopic in the sense that they correctly perceived the benefits of prevention. In reality behavioral biases are likely to affect individual's willingness to undergo screening test and the policy should correct for possible misperception.

All these issues are on our agenda for future research.

[15] *Ex post* the health states is also likely to induce differences in income. While this is not explicitly considered, it is effectively included in our analysis. With our general utility, one can think of the income loss explaining part of the utility cost of disease.

Appendix

A Proof that $\lambda(\theta) \geq 0$

Substituting (37) in (35) yields

$$f(\theta) - \dot{\lambda}(\theta) - \lambda(\theta)\frac{u_{c\theta}}{u_c} - \frac{\mu f(\theta)}{u_c} = 0. \tag{A1}$$

Assume that $\lambda(\theta) < 0$ on some interval $[\theta_a, \theta_b]$. We thus have:

$$\lambda(\theta_a) = \lambda(\theta_b) = 0, \tag{A2}$$

$$\lambda'(\theta_a) \leq 0 \text{ and } \lambda'(\theta_b) \geq 0. \tag{A3}$$

Now consider the dual problem associated to the expenditure minimization in state θ:

$$\begin{aligned} \min_{c(\theta),m(\theta)} \quad & \mathcal{E} = c + m \\ \text{s.t.} \quad & u(c,m,\theta) - \tilde{u} \geq 0 \end{aligned}$$

Denoting by σ the Lagrange multiplier associated to the utility constraint, the first order conditions are

$$1 - \sigma u_c = 0$$

$$1 - \sigma u_m = 0$$

and $u(c,m,\theta) - v(\tilde{u}) = 0$. The solution to this problem yields $c(\tilde{u})$ and $m(\tilde{u})$ and $\mathcal{E}(\tilde{u})$.

Differentiation of the utility constraint yields:

$$u_c\frac{\partial c}{\partial \tilde{u}} + u_m\frac{\partial m}{\partial \tilde{u}} - 1 = 0$$

so that using the envelope theorem

$$\begin{aligned} \frac{\partial \mathcal{E}(\theta,\tilde{u})}{\partial \tilde{u}} &= \frac{\partial c(\theta,\tilde{u})}{\partial \tilde{u}} + \frac{\partial m(\theta,\tilde{u})}{\partial \tilde{u}} \\ &= \frac{1}{u_c} > 0 \end{aligned} \tag{A4}$$

Now combining (A1), (A4) and (A2) yields

$$\frac{\partial \mathcal{E}(\theta, \tilde{u})}{\partial \tilde{u}} = \frac{1}{\mu} - \frac{\dot{\lambda}(\theta)}{\mu f(\theta)} \text{ for } \theta = \{\theta_a, \theta_b\}$$

which using (A3) implies

$$\frac{\partial \mathcal{E}(\theta_a, \tilde{u}(\theta_a))}{\partial \tilde{u}} \geq \frac{\partial \mathcal{E}(\theta_b, \tilde{u}(\theta_b))}{\partial \tilde{u}} \tag{A5}$$

We now show that this inequality implies $\tilde{u}(\theta_b) \leq \tilde{u}(\theta_a)$. Assume instead that $\tilde{u}(\theta_a) < \tilde{u}(\theta_b)$. One has

$$\begin{aligned}\frac{\partial^2 \mathcal{E}(\theta, \tilde{u})}{\partial \tilde{u}^2} &= -\frac{\frac{\partial c(\theta,\tilde{u})}{\partial \tilde{u}} u_{cc} + \frac{\partial m(\theta,\tilde{u})}{\partial \tilde{u}} u_{mm}}{u_{c^2}} \\ &= -\frac{\frac{\partial c(\theta,\tilde{u})}{\partial \tilde{u}} u_{cc} + \left(\frac{1-u_c \frac{\partial c}{\partial \tilde{u}}}{u_m}\right) u_{mm}}{u_{c^2}} \\ &= -\frac{\frac{\partial c(\theta,\tilde{u})}{\partial \tilde{u}} (u_{cc} - 1) + u_{mm}}{u_c^2} > 0\end{aligned}$$

and

$$\frac{\partial^2 \mathcal{E}(\theta, \tilde{u})}{\partial \tilde{u} \partial \theta} = \frac{-u_{c\theta} + u_{m\theta}}{u_c^2} > 0$$

so that $\tilde{u}(\theta_a) < \tilde{u}(\theta_b)$ implies $\partial \mathcal{E}(\theta_a, \tilde{u}(\theta_a)) / \partial \tilde{u} < \partial \mathcal{E}(\theta_b, \tilde{u}(\theta_b)) / \partial \tilde{u}$, which contradicts (A5). Consequently, we must have $\tilde{u}(\theta_b) \geq \tilde{u}(\theta_a)$ but this clearly violates the incentive constraint. Indeed, we have $\theta_b > \theta_a$ and the incentive constraint (22) implies that $\dot{U} < 0$.

Summing up, we have shown that if $\lambda(\theta) < 0$ on some interval $[\theta_a, \theta_b]$, both $\tilde{u}(\theta_a) < \tilde{u}(\theta_b)$ and $\tilde{u}(\theta_b) \geq \tilde{u}(\theta_a)$ are impossible. Consequently, $\lambda(\theta) < 0$ is not possible.

References

[1] Barigozzi, F., (2004), Reimbursing preventive care The Geneva Papers on Risk and Insurance Theory, 29, 165–186.

[2] Besley, T., (1988). Optimal reimbursement health insurance and the theory of Ramsey taxation. Journal of Health Economics, 7(4), 321–336.

[3] Blomqvist, A., (1997), Optimal non-linear health insurance, Journal of Health Economics, 16(3), 303–321.

[4] Cremer, H. and J-M. Lozachmeur, (2022), Coinsurance vs. co-payments: Reimbursement rules for a monopolistic medical product with competitive health insurers, Journal of Health Economics, 84, 102642.

[5] Cremer, H. and P. Pestieau, (1996), Redistributive taxation and social insurance, International Tax and Public Finance, 3, 281–295.

[6] Finkelstein, A., Luttmer, E and M. Notowidigdo, (2013), What good is wealth without health? The effect of health on the marginal utility of consumption, Journal of the European Economic Association 11(1), 221–258.

[7] Eeckhoudt, L., P. Godfroid and M. Marchand, (1998), "Health risk, Preventive Medicine and Curative Medicine," Revue d'economie politique, 108(3), 321–337.

[8] Eeckhoudt, L., P. Godfroid and M. Marchand, (2000), "Co-Payment Rates of Preventive and Curative

[9] Medicine," Revue d'economie politique, 109(4), 483–492.

[10] Ellis, R. and WManning, (2007). Optimal health insurance for prevention and treatment. Journal of Health Economics, 26(6).

[11] Kenkel, D, (2000), Prevention, in: Culyer, A., and J. Newhouse (eds.), Handbook of Health Economics, Elsevier, 1675–1720.

[12] Martinon, P., P. Picard, and A. Raj, (2018), On the design of optimal health insurance contracts under ex post moral hazard, Geneva Risk and Insurance Review, 43, 137–185.

[13] Rochet, JC, (1991), Incentives, Redistribution and Social Insurance. Geneva Risk Insurance Review, 16, 143–165.

[14] Takayama, A., (1985), Mathematical Economics (Second Edition), Cambridge University Press, Cambridge UK.

[15] Zweifel, P., F. Breyer and M. Kifmann, (2009), Health Economics, Second edition, Springer, Berlin