# On the Design of Optimal Health Insurance Contracts under Ex Post Moral Hazard

Pierre Martinon, Pierre Picard, Anasuya Raj

# On the Design of Optimal Health Insurance Contracts under Ex Post Moral Hazard

Pierre Martinon*, Pierre Picard† and Anasuya Raj‡

May 22nd, 2018

**Abstract**

We analyze the design of optimal medical insurance under ex post moral hazard, i.e., when illness severity cannot be observed by insurers and policyholders decide for themselves on their health expenditures. The trade-off between ex ante risk sharing and ex post incentive compatibility is analyzed in an optimal revelation mechanism under hidden information and risk aversion. The optimal contract provides partial insurance at the margin, with a deductible when insurers' rates are affected by a positive loading, and it may also include an upper limit on coverage. The potential to audit the health state leads to an upper limit on out-of-pocket expenses.

*Ecole Polytechnique, Department of Applied Mathematics and INRIA, France. Email: pierre.martinon@polytechnique.edu

†CREST-Ecole Polytechnique, France. Email: pierre.picard@polytechnique.edu. Pierre Picard gratefully acknowledges financial support from LabEX ECODEC.

‡CREST-Ecole Polytechnique, France. Email: anasuya.raj@polytechnique.edu

# 1 Introduction

Ex post moral hazard in medical insurance occurs when insurers do not observe the health states of individuals, and policyholders may exaggerate the severity of their illness - Arrow (1963, 1968), Pauly (1968) and Zeckhauser (1970). Proportional coinsurance under ex post moral hazard (i.e., when insurers pay the same fraction of the health care cost whatever the individuals' expenses) has been considered by many authors, including Zeckhauser (1970), Feldstein (1973), Arrow (1976), Feldstein and Friedman (1977), and Feldman and Dowd (1991). However, while proportional coinsurance has the advantage of mathematical tractability, it is neither an optimal solution to the ex post moral hazard problem, nor an adequate representation of the health insurance policies that we may observe.

To approach this issue in more general terms, we may consider a setting where the policyholder has private information about her illness severity and she chooses her health care expenditures - or equivalently where a provider, acting as a "perfect agent" of the policyholder, prescribes the care that is in the patient's best interest. The contract between insurer and insured specifies the insurance premium and the indemnity schedule, i.e., the indemnity as a (possibly non-linear) function of medical expenses. This is equivalent to a direct revelation mechanism that specifies care expenses and insurance transfers as functions of a message sent by the policyholder about the severity of her illness, and where she truthfully reveals her health state to the insurer. Looking for an optimal non-linear insurance contract under ex post moral hazard is thus equivalent to characterizing the optimal solution to an information revelation problem.

The ex post moral hazard information problem was identified by Zeckhauser (1970), and the corresponding literature is surveyed by Winter (2013). Blomqvist (1997) was the first to address this issue with the modern tools of incentive theory, but he unfortunately overlooked important technical aspects (including bunching and limit

conditions), which considerably reduces the relevance of his conclusions.[1] Ma and Riordan (2002) considered a more specific setting, in which the existence and severity of a disease are private information of the patient, and they showed how the optimal copayment should balance the risk-sharing benefits of greater insurance, against the distortions due to inefficient treatment choices. Drèze and Schokkaert (2013) extended Arrow's theorem of the deductible to the case of ex post moral hazard. However, they directly postulated that the insurance premium is computed with a positive loading factor, presumably because of transaction costs. They did not address the question of whether ex post and ex ante moral hazard differ in this respect, independently of the existence of transaction costs. Our objective is to progress further along these lines, with the double concern of robustness of theoretical conclusions and, as far as possible, conformity with economic reality.

Not surprisingly, as already established by Blomqvist (1997), the trade-off between ex post moral hazard incentives and risk sharing leads to a partial coverage at the margin. However, we will show that, under some assumptions about the probability distribution of health states, it also involves a cap on health expenses and insurance indemnities reached by a non-negligible fraction of policyholders. In other words, the optimal contract specifies a partial reimbursement at the margin, with bunching "at the top".[2] In the terminology of health insurance, such an upper limit on coverage corresponds to a fixed-dollar indemnity plan on a per-period basis, i.e., medical insurance pays at most a predetermined amount over the whole policy year, regardless of the total charges incurred. We will also determine that a deductible is optimal

[1] Blomqvist (1997) argues that the indemnity schedule is $S$-shaped, with marginal coverage increasing for small expenses and decreasing for large expenses. As we will see, this conclusion is not valid when bunching and limit conditions are adequately taken into account.

[2] Bunching may also occur in adverse selection principal-agent models with risk averse agents - Salanié (1990) and Laffont and Rochet (1998) - and in the Mirrlees' optimal income tax model - Lollivier and Rochet (1983), Weymark (1986) and Ebert (1992).

only if insurers charge a positive loading because of transaction costs.[3] Hence, ex post and ex ante moral hazard lead to quite different conclusions about the optimality of deductibles: in the absence of transaction costs, a deductible is optimal under ex ante moral hazard when effort affects the probability of an accident (Holmström, 1979),[4] but not under ex post moral hazard. This characterization is robust to changes in the modelling, including the case where income is affected by a background risk and the case where preferences are not separable between wealth and health.

Partial insurance at the margin and caps on insurance indemnities are frequent, but they are far from being a universal characterization of health insurance, be it offered by social security or by private insurers. In the real world, we also observe limits to out-of-pocket expenses that are usually reached for large inpatient care expenses.[5] This discrepancy between theory and practice may be the consequence of an unrealistic feature of the standard ex post moral hazard model: in practice, patients are not always allowed to choose their health expenses freely. It is a fact that basic health expenses are more or less decided unilaterally by patients, for instance whether they should visit their general practitioners or their dentists to cure benign illnesses, while insurers have control over more serious expenses, in particular surgeries or other types of hospital care.

[3] It is well known that optimal insurance contracts may include a deductible because of transaction costs (Arrow, 1963), ex ante moral hazard (Holmström, 1979) or costly state verification (Townsend, 1979). Drèze and Schokkaert (2013) extend Arrow's theorem of the deductible to the case of ex post moral hazard. Although ceilings on coverage are widespread, they have been justified by arguments that are much more specific: either the insurer's risk aversion for large risks and regulatory constraints (Raviv, 1979), or bankruptcy rules (Huberman et al., 1983) or the auditor's risk aversion in costly state verification models (Picard, 2000).

[4] A straight deductible contract, i.e., full coverage of losses above a deductible, is optimal when effort affects the probability of an accident, but not the probability distribution of losses, conditionally on the occurrence of an accident.

[5] See, for instance, the description of the health insurance plans in the Affordable Care Act at https://www.healthcare.gov/health-plan-information/.

Extending our analysis in that direction, we will immerse the ex post moral hazard problem in a costly state verification setting (Townsend, 1979). There should be no audit for low health expenses, because monitoring the expenses would be cost prohibitive. When health expenses cross a certain threshold, an audit should be triggered, and it is then optimal to provide full coverage at the margin, i.e., to include an out-of-pocket maximum in the indemnity schedule.

In brief, our objective is twofold: firstly, to characterize the optimal health insurance indemnity schedule under ex post moral hazard in a way which is as robust as possible, and, secondly, to extend this analysis to a costly state verification setting. To do so, we will mainly limit ourselves to a simple model, similar to Blomqvist's (1997), with one period, one source of risk and one aggregate medical service, and where the health care providers agency problems are ignored. Needless to say, this is a very restrictive setting, and the literature on health insurance has gone well beyond.[6] Our focus will be limited to the "fundamental trade-off of risk spreading and appropriate incentives" (Cutler and Zeckhauser, 2000, p.576), inherent in the optimal insurance problem under ex post moral hazard, without exploring here these multiple extensions. Obviously, crossing these two perspectives is crucial for reaching a thorough understanding of health insurance markets.

Section 2 introduces our main notations and assumptions. Section 3 characterizes the optimal non-linear insurance contract, when the policyholder's preferences are separable between wealth and health. Theoretical results are derived through optimal control techniques, and they are also solved through a computational approach. Section 4 immerses the ex post moral hazard problem in a costly state verification setting, where health expenses may be audited. Section 5 appraises the robustness of our results by considering alternative models, with correlated background risk, non-

[6] See, in particular, the references provided by Ellis, Jiang and Manning (2015) on multiple health treatment goods, correlated sources of health uncertainty and trade-off between treatment and prevention, and by Pflum (2015) on physician incentives.

separable utility, and insurance loading, respectively. Section 6 briefly investigates the connections between our analysis and public policy issues that are ignored in our analysis, although they are of utmost importance. This includes the redistributive objective of state-driven health insurance regimes, and the inefficiency loss due to the agency relationship between physician and patient. Section 7 concludes. The main proofs are in Appendix 1. Appendix 2 includes details on our computational approach and a complementary set of proofs.

# 2 The model

We consider an individual whose welfare depends both on monetary wealth $R$ and health level $H$, with a bi-variate von Neumann-Morgenstern utility function $U(R, H)$ that is concave and twice continuously differentiable. In the following sections, as in Blomqvist (1997), we restrict attention to the case where $U$ is additively separable between $R$ and $H$, and we will write $U(R, H) = u(R) + H$, with $u' > 0$ and $u'' < 0$. Thus, the individual is income risk averse and illness affects her utility, but it does not affect the marginal utility of income.[7] The non-separability case will be considered in sub-section 5.2. The monetary wealth $R = w - T$ is the difference between initial wealth $w$ and net payments $T$ made or received by the individual for her health care, including insurance transfers.

The health level may be negatively affected by illness, but it increases with health expenditures. This is written as:

$$H = h_0 - \gamma x[1 - v(m)], \gamma > 0,$$

where $h_0$ is the initial health endowment, $x \geq 0$ is the severity of illness (or health state), $m \geq 0$ denotes medical expenses and $\gamma$ is a scaling parameter for the welfare

[7] Regarding the empirical analysis of utility functions that depend on health status, see particularly Viscusi and Evans (1990), Evans and Viscusi (1991), and Finkelstein et al. (2013).

gain from these expenses. We assume that $v(m)$ is concave and twice continuously differentiable, with $v(0) = 0, v'(0) = +\infty, v(m) \in (0,1), v'(m) > 0, v''(m) < 0$ if $m \in (0, M), v'(M) = 0, v(m) = v(M) \leq 1$ if $m \geq M > 0$. Illness severity $x$ is randomly distributed over the interval $[0, a], a > 0$, with c.d.f. $F(x)$ and continuous density $f(x) = F'(x) > 0$ for all $x \in [0, a)$.[8]

# 3 Optimal non-linear insurance

## 3.1 Incentive compatibility

We assume that coverage is offered by risk neutral insurers operating in a competitive market without transaction costs, and that each individual can be insured through only one contract. An insurance contract is characterized by a schedule $I(m)$ that defines the indemnity as a function of health expenditures and by premium $P$. Function $I(.) : \mathbb{R}_+ \to \mathbb{R}_+$ is supposed to be continuous, non-decreasing, piecewise continuously differentiable and such that $I(0) = 0$.[9] We have $T = m + P - I(m)$ and $R = w - T = w - P - m + I(m)$.[10] A type $x$ individual chooses her health care expenses $m(x)$ in

[8] For notational simplicity, we assume that there is no probability weight at the no-sickness state $x = 0$, but the model could easily be extended in that direction.

[9] In addition to being realistic, assuming that $I(m)$ is non-decreasing is not a loss of generality if policyholders can claim insurance payment for only a part of their medical expenses: in that case, only the increasing part of their indemnity schedule would be relevant. Piecewise differentiability means that $I(m)$ has only a finite number of non-differentiability points, which includes the indemnity schedule features that we may have in mind, in particular those with a deductible, a rate of coinsurance or an upper limit on coverage. $I(0) = 0$ corresponds to the way insurance works in practice, but it also acts as a normalization device. Indeed, replacing contract $\{I(m), P\}$ by $\{I(m) + k, P + k\}$ with $k > 0$, would not change the net transfer $I(m) - P$ from insurer to insured, hence an indeterminacy of the optimal solution. This indeterminacy vanishes if we impose $I(0) = 0$.

[10] Our notations are presented by presuming that policyholders pay $m$ (i.e., the total cost of medical services) and they receive the insurance indemnity $I(m)$. However, we may also assume that the

order to maximize her utility, that is

$$m(x) \in \arg\max_{\widetilde{m}\geq 0} \left\{u(w - P - \widetilde{m} + I(\widetilde{m})) + h_0 - \gamma x[1 - v(\widetilde{m})]\right\},$$

and we denote $\widehat{I}(x) \equiv I(m(x))$ the insurance indemnity received by this individual. $I(0) = 0$ implies $m(0) = 0$, and thus we have $\widehat{I}(0) = I(m(0)) = 0$.

The allocation $\{m(x), \widehat{I}(x)\}_{|x\in[0,a]}$ is sustained by a direct revelation mechanism in which health expenditures and the indemnity are respectively $m(\widetilde{x})$ and $\widehat{I}(\widetilde{x})$ when the individual announces that her health state is $\widetilde{x} \in [0, a]$, and where truthfully announcing the health state is an optimal strategy. The characterization of the optimal indemnity schedule $I(.)$ will go through the analysis of the corresponding optimal revelation mechanism $\{m(.), \widehat{I}(.)\}$. Let

$$V(x, \widetilde{x}) = u(w - P + \widehat{I}(\widetilde{x}) - m(\widetilde{x})) + h_0 - \gamma x[1 - v(m(\widetilde{x}))]$$

be the utility of a type $x$ individual who announces $\widetilde{x}$. Thus, incentive compatibility requires

$$x \in \arg\max_{\widetilde{x}\in[0,a]} V(x, \widetilde{x}) \text{ for all } x \in [0, a]. \tag{1}$$

The insurer's break-even condition is written as

$$P \geq \int_0^a \widehat{I}(x) f(x) dx. \tag{2}$$

An optimal revelation mechanism $\{m(.), \widehat{I}(.)\} : [0, a] \to \mathbb{R}^2_+$ maximizes the policyholder's expected utility

$$\int_0^a \left\{u(R(x)) + h_0 - \gamma x[1 - v(m(x))]\right\} f(x) dx, \tag{3}$$

where $R(x) \equiv w - P + \widehat{I}(x) - m(x)$, subject to (1) and (2). Lemma 1 is an intermediary step that will allow us to write this optimization problem in a more tractable way.

insurer and policyholders respectively pay $I(m)$ and $m - I(m)$ to medical service providers. Both interpretations correspond to different institutional arrangements, and both are valid in our analysis.

**Lemma 1** *(i) For any incentive compatible mechanism, $m(x)$ and $\widehat{I}(x)$ are non-decreasing. (ii) There exists a continuous optimal direct revelation mechanism $\{m(.), \widehat{I}(.)\}$. (iii) Any continuous direct revelation mechanism is incentive compatible if and only if*

$$\widehat{I}'(x) = \left[1 - \frac{\gamma x v'(m(x))}{u'(R(x))}\right] m'(x), \tag{4}$$

$$m'(x) \geq 0, \tag{5}$$

*at any differentiability point.*

The monotonicity of incentive compatible mechanisms is intuitive: more severe illnesses induce higher medical expenses and higher insurance compensation. If a revelation mechanism includes discontinuities in $\widehat{I}(x)$ and $m(x)$, then it is possible to reach the same expected utility with lower indemnities and expenses, and such a mechanism would not be optimal. The interpretation of (4) and (5) is as follows. Suppose a type $x$ individual slightly exaggerates the severity of her illness by announcing $\widetilde{x} = x + dx$ instead of $\widetilde{x} = x$. Then, at the first-order, the induced utility variation is $\{u'(R(x))[\widehat{I}'(x) - m'(x)] + \gamma x v'(m(x))m'(x)\}dx$, which cancels out when (4) holds. Monotonicity condition (5) is the local second-order incentive compatibility condition. Symmetrically, it is easy to show that (4)-(5) implies incentive compatibility.

## 3.2 The optimal insurance contract

Let us denote $h(x) \equiv m'(x)$. The optimal revelation mechanism maximizes the policyholder's expected utility given by (3) with respect to $\widehat{I}(x), m(x), h(x), x \in [0, a]$ and $P$, subject to $\widehat{I}(0) = m(0) = 0$, condition (2) and

$$\widehat{I}'(x) = \left[1 - \frac{\gamma x v'(m(x))}{u'(R(x))}\right] h(x), \tag{6}$$

$$m'(x) = h(x), \tag{7}$$

$$h(x) \geq 0 \text{ for all } x, \tag{8}$$

$$\widehat{I}(x) \geq 0 \text{ for all } x, \tag{9}$$

This is an optimal control problem where $\widehat{I}(x)$ and $m(x)$ are state variables and $h(x)$ is a control variable.[11] Propositions 1, 2 and 3 and Corollaries 1 and 2 characterize the optimal solution to this problem as well as the corresponding indemnity schedule $I(m)$.

**Proposition 1** *The optimal mechanism is such that* $0 < \widehat{I}(x) < m(x)$ *for all* $x > 0$.

**Proposition 2** *Assume* $f(x)$ *is non-increasing and* $\ln f(x)$ *is weakly convex. Then there exists* $\overline{x}$ *in* $(0, a]$ *such that*

$$0 < \widehat{I}'(x) < m'(x) \quad \textit{if} \quad 0 < x < \overline{x},$$
$$\widehat{I}(x) = \widehat{I}(\overline{x}), m(x) = m(\overline{x}) \ \textit{if} \ \ \overline{x} < x \leq a.$$

**Corollary 1** $\overline{x} = a$ *if* $x$ *is uniformly distributed over* $[0, a]$.

**Corollary 2** *Assume* $f(a) = f'(a) = 0$, $f''(a) > 0$, *and* $d \ln f(x)/dx$ *and* $d^2 \ln f(x)/dx^2$ *remain finite when* $x \to a$. *Then, we have* $\overline{x} < a$.

Proposition 1 states that the policyholder receives partial but positive compensation in all of the cases where she incurs care expenses. This is an intuitive result, since there is no reason to penalize a policyholder who would announce that her health health expenses are low (i.e., that $x$ is close to 0). However, it sharply contrasts the ex ante moral hazard setting, since we know from Holmström (1979) that, in that case, a

[11] We use Lemma 1-($ii$) to restrict attention to functions $\widehat{I}(x)$ and $m(x)$ that are continuous. Furthermore, $\widehat{I}(x)$ and $m(x)$ are piecewise differentiable because $I(m)$ is piecewise differentiable. This allows us to use Pontryagin's principle in the proof of Proposition 1. In this proof, it is shown that the optimal revelation mechanism is such that $\widehat{I}'(x) \geq 0$. Since $m'(x) \geq 0$, the optimal mechanism will be generated by a non-decreasing indemnity schedule $I(m)$, as we have assumed. Note that Blomqvist (1997) studies a similar optimization problem, but he wrongly ignores the second-order conditions (8) and the sign conditions (9). Nor does he fully consider the technical implications of the assumption $v'(0) = +\infty$, in the absence of which we would have a corner solution with $m(x) = 0$ for $x$ small.

straight deductible may be optimal, and more generally not indemnifying small claims may be part and parcel of an optimal insurance coverage.[12]

The optimal contract trades off risk-sharing and incentives to not overspend for medical services. According to Proposition 2, if $f(x)$ is non-increasing and $\ln f(x)$ is weakly convex,[13] then this trade-off may tip in favor of the incentive effect when $x$ is large enough. If $x$ is lower than $\overline{x}$, then $m(x)$ and $\widehat{I}(x)$ monotonically increase, with an increase in the out-of-pocket expenses $m(x)-\widehat{I}(x)$, when $x$ goes from 0 to $\overline{x}$. When $x \geq \overline{x}$, there are ceilings $m(\overline{x})$ and $\widehat{I}(\overline{x})$, respectively, for expenses and indemnity. Corollaries 1 and 2 illustrate the two possible cases $\overline{x} = a$ (no bunching) and $\overline{x} < a$ (bunching), respectively. There is no bunching when the illness severity is uniformly distributed in the $[0, a]$ interval. If the density function of $x$ decreases to zero when $x$ goes to $a$ and is differentiable at $x = a$, then Corollary 2 provides a sufficient condition for bunching to be optimal. In the first case, the probability of the highest severity levels remains large enough for the capping of expenditures and indemnities to be sub-optimal, while in the second case it is optimal. If we consider the differentiability of density $f(x)$ at the top as a natural assumption, then Corollary 2 provides support for upper limits in optimal insurance indemnity schedules.

In what follows, we provide detailed intuition for the possibility of bunching, and particularly for the reason why it occurs under the assumptions of Corollary 2. We may first observe that increasing $\widehat{I}(x)$ is a way to incentivize type $x$ policyholders to report her health state truthfully (i.e., not to report $\widetilde{x} > x$) and also to improve her coverage. However, as highlighted in Lemma 1-*iii*, this can be done only by increasing $m(x)$ in order to preserve the incentives of type $x'$ policyholders for $x' < x$. This increase in $m(x)$ will exacerbate the overexpense problem. Bunching occurs when the

[12] Note the relationship of Proposition 1 with optimal insurance under (ex ante) moral hazard when effort affects the distribution of losses should an accident occur, but not the probability of the accident itself. In that case, it may be optimal to fully cover small losses without a deductible. See Rees and Wambach (2008).

[13] This is the case, for instance, if the distribution of $x$ is uniform or exponential.

negative effect of an increase in health care expenses outweighs the positive effect of a more complete insurance coverage.

In order to understand this trade-off more completely, let us consider the co-state variables $\mu_1(x)$ and $\mu_2(x)$, associated with $\widehat{I}(x)$ and $m(x)$, respectively. The evolution laws of $\mu_1(x)$ and $\mu_2(x)$ are derived from optimal control theory, and they are used extensively in the proofs. They correspond to the first-order variations in the objective of the partial optimization problem limited to $[x, a]$, following discontinuous small variations $\Delta\widehat{I}(x) > 0, \Delta m(x) > 0$.[14] A discontinuous increase in $\widehat{I}(x)$ would be advantageous because it improves risk coverage and it corresponds to a relaxation of the upward incentive compatibility constraint (type $x$ individuals have less incentive to report $\widetilde{x}$ larger than $x$). Conversely, an upward discontinuous shift in $m(x)$ would exacerbate the distortion between the marginal utility of wealth $u'(R(x))$ and the marginal utility of health expenses $\gamma x v'(m(x))$. It is therefore intuitive that $\mu_1(x) > 0, \mu_2(x) < 0$, which is established and used in the proofs, as well as the transversality conditions $\mu_1(a) = \mu_2(a) = 0$.

Lemma 1 shows that we should have

$$\frac{\Delta\widehat{I}(x)}{\Delta m(x)} = 1 - \frac{\gamma x v'(m(x))}{u'(R(x))}$$

for such discontinuous upward variations to be approximated, as closely as we would like, by incentive compatible continuous trajectories $\widehat{I}(x)$ and $m(x)$. Keeping in mind this link between feasible variations in $\widehat{I}(x)$ and $m(x)$, let us denote

$$\varphi(x) \equiv \mu_1(x)\left[1 - \frac{\gamma x v'(m(x))}{u'(R(x))}\right] + \mu_2(x).$$

Function $\varphi(x)$ sums up the negative effect of an increase in $m(x)$ and the positive effect of the induced increase in $\widehat{I}(x)$, weighted by $\mu_2(x)$ and $\mu_1(x)$, respectively, with

[14] In more technical terms, we may define the value function $v(I_0, m_0, x)$ to be the greatest expected utility over $[x, a]$, with unchanged insurance expected cost, if we start at $\widehat{I}(x) = I_0, m(x) = m_0$. The vector of costates $(\mu_1(x), \mu_2(x))$ is the gradient at $x$ of the value function, evaluated along the optimal trajectory.

$\varphi(a) = 0$. The previous intuitive reasoning suggests (and the proof confirms) that an optimal solution should satisfy $\varphi(x) = 0$ if $h(x) > 0$ and $\varphi(x) \leq 0$ if $h(x) = 0$.[15] In particular, if $m'(x) = h(x) > 0$ we have $\gamma x v'(m(x)) < u'(R(x))$ and thus $\widehat{I}'(x) > 0$, which corresponds to the two possible regimes described in Proposition 2: $m(x)$ and $\widehat{I}(x)$ are simultaneously increasing or stationary. Bunching occurs when the negative effect of an increase in health expenses outweighs the gains from an increase in insurance coverage. More details are provided in the following remark.

**Remark 1** *To be more explicit about the conditions under which there is bunching, let us assume $\overline{x} < a$, with $m(x) = \overline{m}$ and $R(x) = \overline{R}$ when $x \in [\overline{x}, a]$. Then, it can be shown that*

$$\frac{\mu_1(x)}{1 - F(x)} = u'(\overline{R}) - \lambda,$$
$$\frac{\mu_2(x)}{1 - F(x)} = -u'(\overline{R}) + \gamma v'(\overline{m}) \int_x^a t \frac{f(t)}{1 - F(x)} dt,$$

*if $x \in [\overline{x}, a]$, where $\lambda$ is the (positive) Lagrange multiplier associated with the insurer's break-even constraint (2).*[16] *Intuitively, when there is bunching, the trajectory $m(x), \widehat{I}(x)$ is stationary, and the first-order effect of an increase $\Delta\widehat{I}(x)$ on the policyholder's expected utility is just the difference between the policyholder's marginal utility gain $u'(\overline{R})\Delta\widehat{I}(x)$ and the marginal loss due to the induced increase in insurance cost $\lambda\Delta\widehat{I}(x)$, multiplied by the probability $1 - F(x)$ of being in $[x, a]$. Similarly, the first-order effect of an increase $\Delta m(x)$ can be approximated by the variation of the policyholder's surplus $-[u'(\overline{R}) - \gamma t v'(\overline{m})]\Delta m(x)$ averaged over $[x, a]$ according to the conditional density $f(t)/[1 - F(x)]$ . Hence, for all $x$ in $[\overline{x}, a]$, we have*

$$\varphi(x) = G(x)[1 - F(x)],$$

[15] $\varphi(x)$ is called a "switching function" in the optimal control terminology, because its sign determines the sign of the control.

[16] These conditions can be deduced from the trajectories of $\mu_1(x)$ and $\mu_2(x)$.

*where*

$$G(x) = -\lambda \left[1 - \frac{\gamma x v'(\overline{m})}{u'(\overline{R})}\right] + \gamma v'(\overline{m}) \left[\int_x^a t \frac{f(t)}{1-F(x)} dt - x\right].$$

*When $x$ is uniformly distributed, we have $f(x) = 1/a, F(x) = x/a$ and $\varphi(x)$ is a second degree polynomial when $x \in [\overline{x}, a]$. To simplify things, assume that the control $h(x)$ is continuous at $x = \overline{x}$.*[17] *Then, the switching function $\varphi(x)$ is differentiable at $x = \overline{x}$, with $\varphi(\overline{x}) = \varphi'(\overline{x}) = 0$, which is incompatible with $\varphi(a) = 0$ when $\varphi(x)$ is a second degree polynomial. Hence, bunching cannot occur in that case, as established in Corollary 1.*[18] *Under the assumptions of Corollary 2, we have $\varphi(a) = \varphi'(a) = \varphi''(a) = 0$ and $\varphi'''(a) = -f''(a)G(a)$. Hence $G(a) < 0$ is a sufficient condition for $\varphi''(x) < 0$ when $x$ is close to $a, x < a$. In that case, the switching function $\varphi(x)$ has a local maximum at $x = a$, with $\varphi(x) < 0$ when $x$ is close to $a$. The proof of Corollary 2 shows that this is actually what occurs.*

**Remark 2** *Proposition 2 is based on assumptions that we may find overly restrictive. It can be reformulated in a weaker form, by only assuming that $f(x)$ is non-increasing and $\ln f(x)$ is weakly convex in a subinterval $[x_0, x_1] \subset [0, a]$, and in that case there exists $\overline{x} \in (x_0, x_1]$ such that $\widehat{I}(x)$ and $m(x)$ are increasing over $[x_0, \overline{x})$ and constant over $[\overline{x}, x_1]$. For instance, if $x$ is log-normal, with $a = +\infty, \mathbb{E}[\ln(x)] = \mu$ and $Var[\ln(x)] = \sigma^2$, then $\ln f(x)$ is decreasing and convex when $x \geq \exp(1 + \mu - \sigma^2) \equiv x_0$. In that case, $\widehat{I}(x)$ and $m(x)$ are increasing in $[x_0, \overline{x})$, and constant in $[\overline{x}, +\infty)$, with $\overline{x} > x_0$. Similarly, the proof of Corollary 2 shows that bunching at the top is optimal without using the assumptions made in Proposition 2. In other words, these assumptions guarantee that there exists a threshold $\overline{x}$ such that bunching occurs if and only if $x \geq \overline{x}$, but they are not required to show that there is bunching when $x$ is large enough.*

**Proposition 3** *Under the assumptions of Proposition 2, the optimal indemnity sched-*

[17] The proofs do not require this assumption.

[18] A similar but more complex argument is used in the proof of Proposition 2 to show that bunching cannot occur in intervals interior to $[0, a]$.

*ule $I(m)$ is such that*

$$I'(m) \in (0,1) \quad \textit{if} \quad m \in (0, \overline{m}),$$
$$I'(\overline{m})_- = 0 \ \textit{if} \ \overline{x} = a, I'(\overline{m})_- > 0 \ \textit{if} \ \overline{x} < a,$$
$$I(m) = I(\overline{m}) \quad \textit{if} \quad m \geq \overline{m},$$

*where $\overline{m} = m(\overline{x})$. We have $I'(0) \geq 0$ and $\lim_{m \to 0} - mv''(m)/v'(m) < 1$ is a sufficient condition for $I'(0) > 0$.*

The characterization of the indemnity schedule $I(m)$ provided in Proposition 3 is derived from $I(m(x)) \equiv \widehat{I}(x)$ , which gives

$$I'(m) = \frac{\widehat{I}'(x)}{m'(x)} = 1 - \frac{\gamma x v'(m(x))}{u'(R(x))} < 1,$$

if $m = m(x)$ and $0 < x < \overline{x}$. If there is no bunching, then there is no distortion at the top, i.e., the marginal benefit drawn from health care expenses is equal to the marginal utility of wealth: this corresponds to $u'(\overline{R}) - \gamma \overline{x} v'(\overline{m}) = 0$, and thus $I'(\overline{m}) = 0$. We have $I'(\overline{m})_- > 0$ in the case of bunching.

Hence, the indemnity schedule has a slope between 0 and 1 in its increasing part. At the bottom, there is no deductible, contrary to case of ex ante moral hazard. At the top, in the case of bunching, the indemnity schedule has an angular point at $m = \overline{m}$, and all the individuals with an illness severity larger than $\overline{x}$ are bunched with the same amounts of health expenses $\overline{m}$ and insurance indemnity $I(\overline{m})$.[19] In the absence of bunching, the population of policyholders is spread from $m(0) = \widehat{I}(0) = 0$ to $m(a) > \widehat{I}(a) > 0$ when $x$ increases from 0 to $a$, with different choices for different illness severity levels. The slope of the indemnity schedule $I(m)$ goes to zero when $m$ increases to $\overline{m} = m(a)$ because $\gamma a v'(\overline{m}) = u'(\overline{R})$, with $\overline{R} = R(a)$. This corresponds

[19] In practice, the optimal policy could be approximated by a piecewise linear schedule with slope between 0 and 1 until the upper limit $\overline{m}$ and with a capped indemnity when $m > \overline{m}$. It would be interesting to estimate the welfare loss associated with this piecewise linearization. The simulations presented in Section 3.3 suggest that this loss may be low.

to the absence of distortion at the top of the interval $[0, a]$ when there is no bunching, a property shared by other principal-agent models with hidden information and risk-averse agent, such as Salanié (1990) and Laffont-Rochet (1998).

Propositions 2 and 3 justify the existence of a cap on indemnity $I(\overline{m})$, but they also show that medical expenses should not increase in illness severity after the reimbursement ceiling is reached. Intuitively, if $\widehat{I}(x)$ is constant and $m(x)$ increases when $x$ is large, then slightly perturbing the trajectory $\widehat{I}(x)$ so that it is monotonically increasing, with a compensating increase in premium $P$, would improve risk sharing while preserving incentive compatibility. In other words, the profiles of medical expenses and insurance indemnities move simultaneously, and placing a ceiling on insurance indemnities only makes sense because medical expenses are also capped.[20]

## 3.3 Simulations

Simulations are performed by transforming the infinite dimensional optimal control problem into a finite dimensional optimization problem, through a discretization of $x$, applied to the state and control variables, as well as the dynamic equations.[21] We assume that $x$ is distributed over $[0, 10]$ (that is, $a = 10$), either exponentially, i.e., $f(x) = \lambda e^{-\lambda x} + e^{-\lambda a}/a$, with $\lambda = 0.25$,[22] or uniformly, i.e., $f(x) = 1/a$. We also assume $v(m) = \sqrt{m}/[1 + \sqrt{m}]$, with $\gamma = 0.2$ and utility is CARA: $u(R) = -e^{-sR}$, with $s = 10$. The numerical solver leads to optimal functions $\widehat{I}(x)$ and $m(x)$ - and also to $h(x)$ and $P$ - and thus to function $I(m)$ through $I(m(x)) = \widehat{I}(x)$ for all $x \in [0, a]$.

Figure 1 represents the optimal indemnity schedule $I(m)$ and indifference curves in

[20] The same intuition is at work to show that $\widehat{I}'(x) > 0$ when $x$ is close to zero, and thus that the indemnity schedule should not include a deductible, with additional technical specificities induced by the sign constraint $\widehat{I}(x) \geq 0$.

[21] We use the Bocop software (see Bonnans et al., 2016, and http://bocop.org). We refer the reader to Appendix 2-A and, for instance, to Betts (2001) and Nocedal and Wright (1999) for more details on direct transcription methods and non-linear programming algorithms.

[22] Note that $f(a)$ and $f'(a)$ are close to 0 when $a$ is large.

the $(m, I)$ space for $x \in \{0.3, 7, 9\}$ when $x$ is uniformly distributed. Parameters $\sigma$ and $k$ will be introduced later: they correspond to a loading factor and to the intensity of a background risk, respectively. Here, both are equal to 0, since there is no loading and no background risk. The optimal type $x$ indifference curve is tangent to the indemnity schedule for expenses $m(x)$. As stated in Corollary 1, there is no bunching: $m(x)$ goes from $m(0) = 0$ to $\overline{m} = m(10) \simeq 0.7002$ and $\widehat{I}(x) = I(m(x))$ goes from $I(0) = 0$ to $I(0.7002) \simeq 0.3863$, when $x$ goes from 0 to 10. There is no deductible (i.e. $I'(0) > 0$) and the marginal coverage cancels at the top, that is $I'(0.7002) = 0$. The locus of function $I(m)$ is completed by a flat part for $m > \overline{m}$, while preserving differentiability.

The slope of the type $x$ policyholder indifference curve is written as

$$\frac{dI}{dm}_{|EU=const.} = \frac{U'(w - P - m + I) - \gamma x v'(m)}{U'(w - P - m + I)},$$

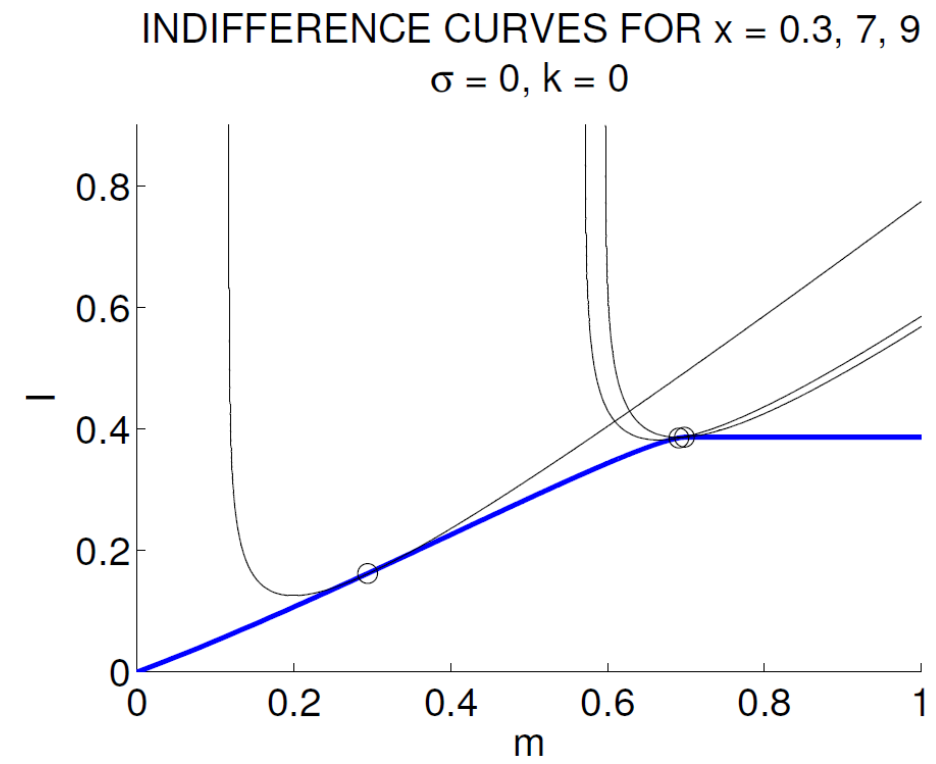
and it cancels at the top of the increasing part of the $I(m)$ curve, when $m = \overline{m}$, which corresponds to the optimal policyholder's choice when $x = 10$. The optimal choices of the policyholder are spread from $m = 0$ to $m = \overline{m}$ when $x$ goes from 0 to 10, and the flat part of the $I(m)$ curve is never reached.

Figure 2 corresponds to the case of an exponential distribution, with indifference curves also drawn for $x \in \{0.3, 7, 9\}$. Now, there is bunching at the top, as expected from Corollary 2. We have $\overline{x} \simeq 6.7$ and $\overline{m} \simeq 0.7490$. $I(m)$ has an angular point at $m = \overline{m}$, with $I(\overline{m}) \simeq 0.4253$. Figure 2 illustrates the case of types $x = 7$ and $x = 9$: in both cases, the optimal expenses are equal to $\overline{m}$. As in Figure 1, we have $I'(0) > 0$.

**Figures 1 and 2**

# 4 Auditing

We still consider allocations $\{m(x), \widehat{I}(x)\}_{|x \in [0,a]}$ that are induced by non-linear indemnity schemes $I(m)$ with $\widehat{I}(x) \equiv I(m(x))$. However, as in the costly state verification



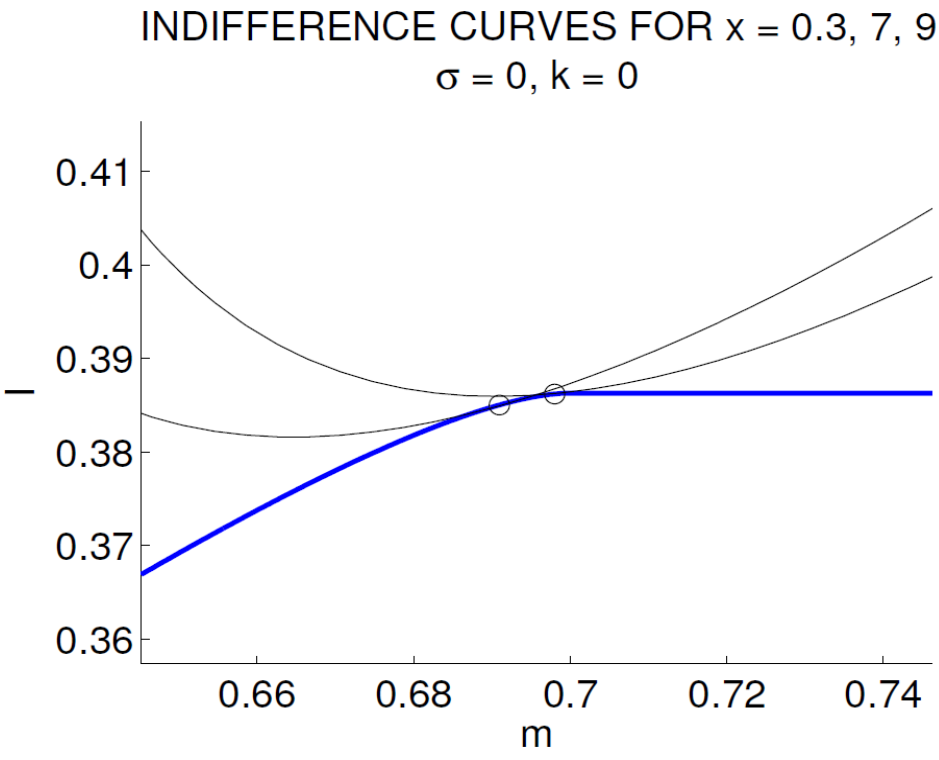
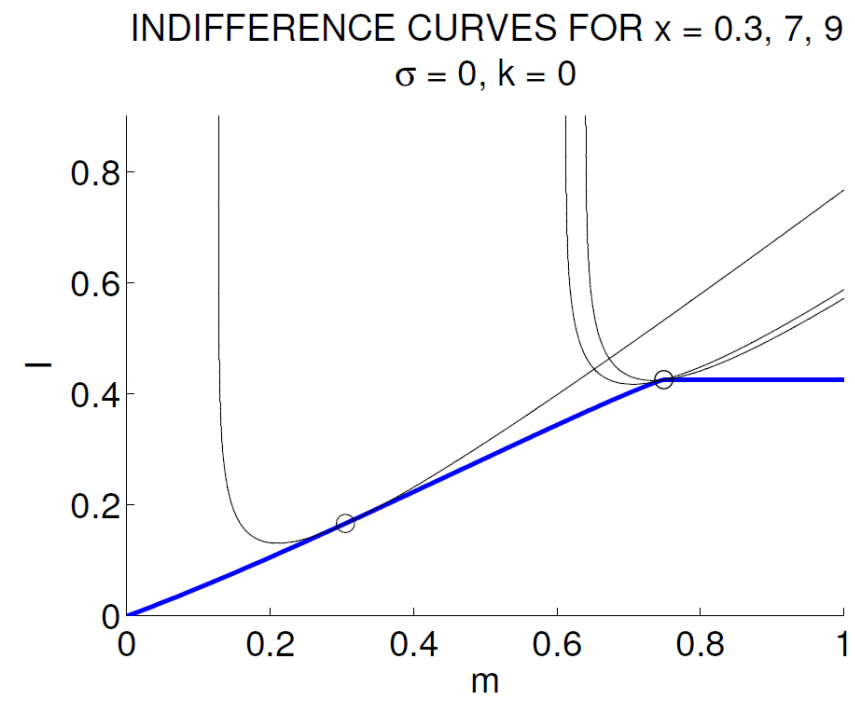




**Figure 1**

**Uniform distribution – No bunching**



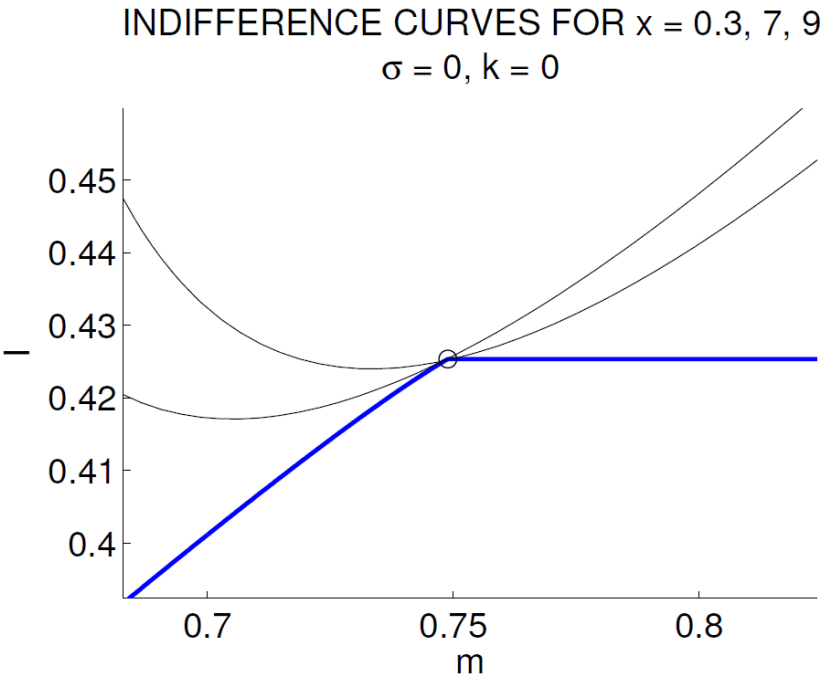




**Figure 2**

**Exponential distribution - Bunching**

approach introduced by Townsend (1979), we now assume that the insurer can verify the health state $x$ by incurring an audit cost $c > 0$. We restrict attention to a deterministic auditing strategy, in which the insurer audits the insurance claims larger than a threshold $m^*$, or equivalently, since $m(x)$ will be non-decreasing, when $x > x^*$ $= \inf\{x : m(x) > m^*\}$.[23] In the case of an audit, the policyholder's medical expenses are capped by the expense profile $m(x)$.[24] In other words, audit allows the insurer to monitor the policyholder's medical expenses. Thus, a type $x$ individual chooses her health expenses $m'$ under the constraint $m' \leq \sup\{m^*, m(x)\}$, and she receives indemnity $I(m')$.

**Definition 1** $\{I(m), m(x), m^*, P\}_{|x\in[0,a]}$ *implements the allocation* $\{m(x), \widehat{I}(x), x^*, P\}_{|x\in[0,a]}$ *if* $(i) : m(x)$ *is an optimal expense choice of type* $x$ *individuals under indemnity schedule* $I(m)$*, constraint* $m \leq \sup\{m^*, m(x)\}$*, and insurance premium* $P$*,*$(ii) : \widehat{I}(x) = I(m(x))$ *for all* $x \in [0, a]$*, and* $(iii)$ *: there is audit when* $x > x^* = \inf\{x : m(x) > m^*\}$*.*

For the sake of realism, we restrict attention to (piecewise differentiable) continuous functions $I(m)$ such that $I'(m) \leq 1$ if $m \geq m^*$, although, as we will see, an upward discontinuity of $I(m)$ at $m = m^*$ would be optimal.[25] We denote $g(x) \equiv \widehat{I}'(x)$ when

[23] More generally, the insurer could randomly audit claims, the probability of triggering an audit depending on the size of the claim. See the references in Picard (2013) on deterministic and random auditing for insurance claims.

[24] The policyholder is subject to prior authorization for increasing her medical expenses above $m^*$. After auditing the health state, this authorization will be granted but capped by $m(x)$ if $x > x^*$, and otherwise it will be denied.

[25] Since an upward discontinuity of $I(m)$ at $m = m^*$ dominates the optimal solution when $I(m)$ is constrained to be continuous, increasing $I(m)$ as much as possible in a small interval $(m^*, m^* + \varepsilon)$ would bring the continuous function $I(m)$ arbitrarily close to this discontinuous function. No optimal solution would exist in the set of continuous functions $I(m)$. Thus, in addition to being realistic from an empirical point of view, the assumption $I'(m) \leq 1$ if $m \geq m^*$ eliminates this reason for which an optimal solution may not exist. As previously shown, we have $I'(m) < 1$ in the no-audit regime where $m < m^*$.

$x > x^*$ , and, as previously, $h(x) = m'(x)$ for all $x$. The optimization problem is written as

$$\max \int_0^a \left\{ u(w - P + \widehat{I}(x) - m(x)) + h_0 - \gamma x[1 - v(m(x))] \right\} f(x)dx$$

with respect to $\widehat{I}(x), m(x), g(x), h(x), x^* \in [0, a]$, and $P$, subject to $\widehat{I}(0) = 0$,(7) and (9) for all $x$, (6) and (8) if $x \leq x^*$, and

$$\widehat{I}'(x) = g(x) \text{ if } x > x^*, \tag{10}$$

$$0 \leq g(x) \leq h(x) \text{ if } x > x^*, \tag{11}$$

$$P = \left[ \int_0^{x^*} \widehat{I}(x) f(x)dx + \int_{x^*}^a [\widehat{I}(x) + c] f(x)dx \right]. \tag{12}$$

Condition (11) follows directly from $0 \leq I'(m) \leq 1$ when $m \geq m^*$ since $I'(m(x)) = \widehat{I}'(x)/m'(x) = g(x)/h(x)$. Now, we have an optimal control problem with two regimes, according to whether $x$ is smaller or larger than $x^*$ and where $g(x)$ is a new control variable when $x > x^*$.[26] In the first stage, we will characterize the optimal trajectory $\widehat{I}(x), m(x)$ over the interval $(x^*, a]$, for a given trajectory $\widehat{I}(x), m(x)$ over $[0, x^*]$ and for given values of $P$ and $x^*$. In the second stage, we will solve for the optimal trajectory $\widehat{I}(x), m(x), x \in [0, x^*]$ and for the optimal values of $P$ and $x^*$, given the characterization of the optimal continuation trajectory.

Let $I^* = \widehat{I}(x^*)$ and $m^* = m(x^*)$, with $I^* \leq m^*$. For $\{\widehat{I}(x), m(x), x \in [0, x^*]\}, P$ and $x^*$ given and such that

$$P \geq \int_0^{x^*} \widehat{I}(x) f(x)dx + (I^* + c)[1 - F(x^*)], \tag{13}$$

$$u'(w - P - m^* + I^*) \geq \gamma x^* v'(m^*). \tag{14}$$

$\{\widehat{I}(x), m(x), g(x), h(x),\ x \in (x^*, a]\}$ maximizes

$$\int_{x^*}^a \left\{ u(w - P + \widehat{I}(x) - m(x)) + h_0 - \gamma x[1 - v(m(x))] \right\} f(x)dx, \tag{15}$$

[26] If $c = 0$, then the first-best allocation would be feasible with $x^* = 0$, that is by auditing the health state in all possible cases. Thus, choosing $x^*$ smaller than $a$ is optimal when $c$ is not too large, and this is what we assume in what follows.

subject to (7), (10)-(12). This is a subproblem restricted to $x \in (x^*, a]$. Note that $\widehat{I}(x) = I^*, m(x) = m^*, g(x) = 0, h(x) = 0$ for all $x \in (x^*, a]$ is a feasible solution to this subproblem because of (13). Conversely, (13) holds for any solution such that $g(x) = \widehat{I}'(x) \geq 0$ for all $x \in (x^*, a]$. Furthermore, we have

$$u'(w - P - m(x) + I(m(x))[1 - I'(m(x))] - \gamma x v'(m(x)) = 0,$$

for all $x \leq x^*$. Using $I'(m) \leq 1$ for all $m$ gives:

$$u'(w - P - m(x) + I(m(x)) - \gamma x v'(m(x)) \geq 0,$$

and using $m^* = m(x^*)$ implies (14). Thus, we may characterize the optimal solution to this subproblem by assuming (13) and (14) without further loss of generality.

**Lemma 2** *When $x^*, m^*, P$ and $I^*$ satisfy (13) and (14), the optimal continuation allocation is such that*

$$\begin{aligned}
&\widehat{I}'(x) = m'(x) = 0 \text{ if } x \in [x^*, \widetilde{x}], \\
&\widehat{I}'(x) = 0, m'(x) = -\frac{\gamma v'(m(x))}{\gamma x v''(m(x)) + u''(R(x))} \text{ if } x \in [\widetilde{x}, \widehat{x}], \\
&\widehat{I}'(x) = m'(x) = -\frac{v'(m(x))}{x v''(m(x))} \text{ if } x \in (\widehat{x}, a],
\end{aligned}$$

*where $R(x) = w - P - m(x) + I^*$ and $x^* \leq \widetilde{x} \leq \widehat{x} < a$, with $x^* = \widehat{x}$ for the optimal allocation.*

Lemma 2 characterizes an optimal continuation allocation, with $x^*, m^*, P$ and $I^*$ considered as parameters. In particular, $I^*$ and $m^*$ may differ from the optimal solutions $\widehat{I}(x^*)$ and $m(x^*)$. If the hypothesized values of $I^*$ and/or $m^*$ are large (in particular, if they are larger than their optimal value around $x^*$ in the global problem), then an optimal solution of the restricted problem may consist in keeping $\widehat{I}(x)$ and/or $m(x)$ constant when $x$ larger than but close to $x^*$, and to increase $\widehat{I}(x)$ and $m(x)$ only when $x$ is substantially larger than $x^*$. Lemma 2 says that the increase

in $\widehat{I}(x)$ should be concentrated on the highest values of $x$, that is when $x > \widehat{x}$ with $\widehat{x} \in [x^*, a]$: these values correspond to the largest health expenses, and thus to the cases where the marginal utility of wealth is the largest. In the lowest part of the interval, i.e., when $x < \widetilde{x}$, not increasing health expenses may be optimal. Lemma 2 also states that the optimal insurance contract provides full coverage at the margin, that is $\widehat{I}'(x) = m'(x)$, when $x > \widehat{x}$. There is nothing astonishing here: in the case of an audit, there is no more asymmetry of information, and the policyholder should be fully compensated for any increase in her insurable losses.[27]

Lemma 2 states that three regimes may potentially exist in the restricted problem: $\widehat{I}'(x) = m'(x) = 0$ when $x^* < x \leq \widetilde{x}, \widehat{I}'(x) = 0, m'(x) > 0$ when $\widetilde{x} < x \leq \widehat{x}$, and $\widehat{I}'(x) > 0, m'(x) > 0$ when $\widehat{x} < x < a$. However, if the two first regimes were part and parcel of the globally optimal solution, i.e., if $x^* < \widetilde{x}$ and/or $\widetilde{x} < \widehat{x}$, then a costly audit woud be performed when $x \in (x^*, \widehat{x}]$, although the same insurance indemnity $I^*$ is paid when the policyholder chooses $m \in (m^*, m(\widehat{x}))$ than when she choses $m^*$. This would be obviously suboptimal. In other words, a globally optimal allocation should be such that $x^* = \widehat{x}$, because auditing is useless if the indemnity does not increase above the maximum $I^*$ that can be reached in the no-audit regime.

Let $V(m^*, I^*, x^*, P, A)$ be the value of the integral (15) at an optimal continuation equilibrium, where

$$A = \int_0^{x^*} \widehat{I}(x) f(x) dx. \tag{16}$$

[27] See Gollier (1987) and Bond and Crocker (1997) for similar results; see also Picard (2013) for a survey on deterministic auditing in insurance fraud models. Lemma 2 also characterizes the optimal health expenses profile $m(x)$ when there is auditing and full insurance at the margin (that is when $x > \widehat{x}$): we have $m'(x) = -v'(m(x))/xv''(m(x))$, which means that the increase in health expenses which follows a unit increase in the illness severity $x$ is equal to the inverse of the elasticity of the marginal efficiency of health expenses $v'(m(x))$. Equivalently, the marginal utility of health care expenses $\gamma x v'(m(x))$ should remain constant in the auditing regime.

Our global optimization problem can be rewritten as

$$\max \int_0^{x^*} \left\{ u(w - P + \widehat{I}(x) - m(x)) + h_0 - \gamma x[1 - v(m(x))] \right\} f(x)dx + V(m^*, I^*, x^*, P, A)$$

with respect to $\{\widehat{I}(x), m(x), g(x), h(x), x \in [0, x^*]\}, x^* \geq 0, A$ and $P$, subject to $\widehat{I}(0) = 0, I^* = \widehat{I}(x^*), m^* = m(x^*)$, (6)-(9) and (16). The optimal solution to this problem and the corresponding indemnity schedules are characterized as follows.

**Proposition 4** *The optimal mechanism with audit is such that $x^* > 0$, with*

$$\widehat{I}'(x) = m'(x) > 0 \quad \textit{if} \quad x \in (x^*, a],$$

*and with an upward discontinuity of $\widehat{I}(x)$ and $m(x)$ at $x = x^*$. Furthermore, under the same assumptions as Proposition 2, there is $\overline{x} \in (0, x^*]$ such that*

$$0 < \widehat{I}'(x) < m'(x) \quad \textit{if} \quad 0 \leq x < \overline{x},$$
$$\widehat{I}(x) = \widehat{I}(\overline{x}), m(x) = m(\overline{x}) \ \textit{if} \ \overline{x} < x \leq x^*.$$

**Proposition 5** *Under the same assumptions as Proposition 2, the optimal indemnity schedule with audit is such that $m^* = \overline{m} \equiv m(\overline{x}) > 0$, and*

$$I'(m) \in (0, 1) \quad \textit{if} \quad m \in (0, \overline{m}),$$
$$I'(m) = 1 \quad \textit{if} \quad m > \overline{m}.$$

Propositions 4 and 5 show that auditing allows the insurer to offer a protective shield that limits the policyholder's copayment $m(x) - \widehat{I}(x)$. This copayment increases with the expenses when there is no audit, and it reaches an out-of-pocket maximum $\overline{m} - I(\overline{m})$ when the expenses reaches the threshold $m^* = \overline{m} \equiv m(\overline{x})$ above which an audit is triggered. The threshold $\overline{m}$ is reached by a positive mass subset of individuals (those with $x \in [\overline{x}, x^*]$) in the case of bunching. The incentive compatibility constraint vanishes when the health state is audited, which explains why crossing the border between the two regimes should be accompanied by an upward jump in health expenses from $\overline{m}$ to $m(x^*)$, and insurance payment from $I(\overline{m})$ to $I(m(x^*)) = I(\overline{m}) + m(x^*) - \overline{m}$.[28]

[28] Of course, this discontinuity of function $m(x)$ at $x = x^*$ is compatible with a continuous function $I(m)$.

Proposition 5 is illustrated in Figure 3, in the exponential distribution case with $c = 0.25$. We have $x^* \simeq 4.95$. There is coinsurance at the margin, with bunching at the top when $m < m^* = \overline{m}$, and an upward discontinuity of $\widehat{I}(x)$ and $m(x)$ at $x = x^*$. There is full insurance at the margin, that is $I'(m) = 1$ when $m \geq m^*$, with a limit of out-of-pocket expenses equal to $m^* - I(m^*)$. In Figure 3-bottom, the two regimes of the $I(m)$ locus are patched together by a dotted line from $m^* = \overline{m} \simeq 0.41$ to $m(x^*) \simeq 0.95$ with constant slope equal to one, in order to define $I(m)$ for all $m \geq 0$, but $m$ is never chosen in $(\overline{m}, m(x^*))$.[29]

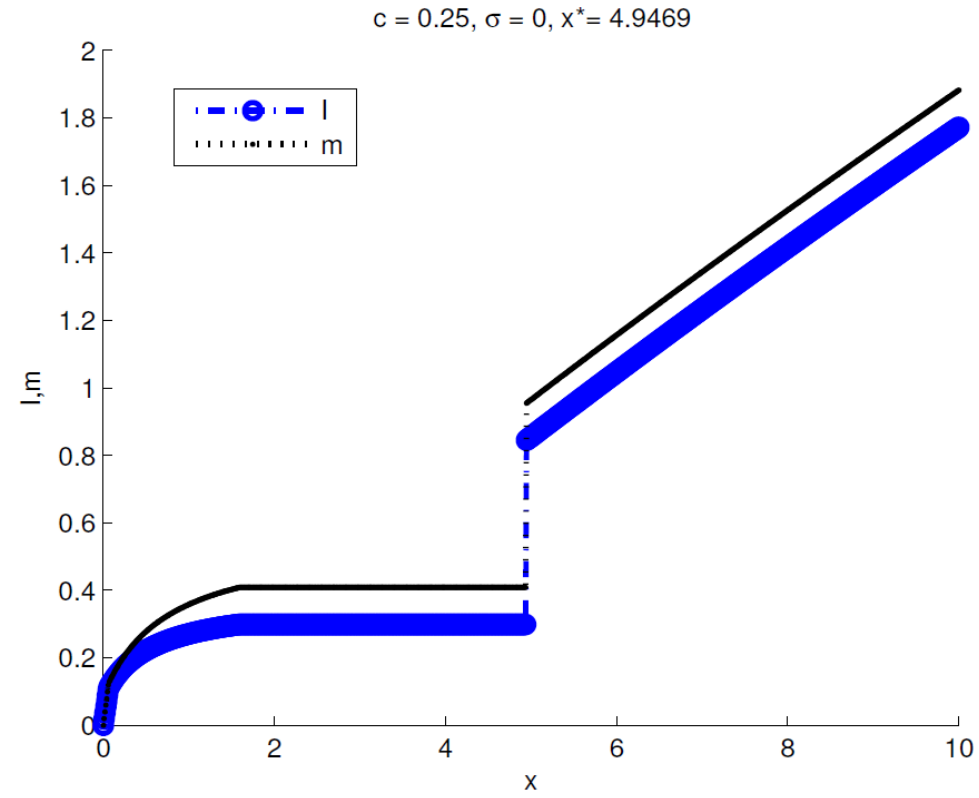
The dependency between the threshold $x^*$ and the audit cost $c$ is simulated in Figure 4. As expected, the larger the audit cost, the larger the threshold above which it is optimal to audit health care expenses.

**Figures 3 and 4**

# 5 Alternative models and robustness

The previous sections have shown how optimal insurance under ex post moral hazard involves partial coverage at the margin, no deductible and, possibly, an upper limit on medical care expenses and coverage, as long as an audit of the patient's health state is not required. This section exlores the robustness of these conclusions by considering alternative modeling options. For the sake of brevity, we will limit ourselves here to the most simple model of sections 2 and 3, without auditing.

[29] The bunching of types is no longer sustained by a kink in the indemnity schedule $I(m)$ at $m = \overline{m}$, but by the threat of an audit, since increasing expenses above $\overline{m}$ will not be possible if $x \leq x^*$.



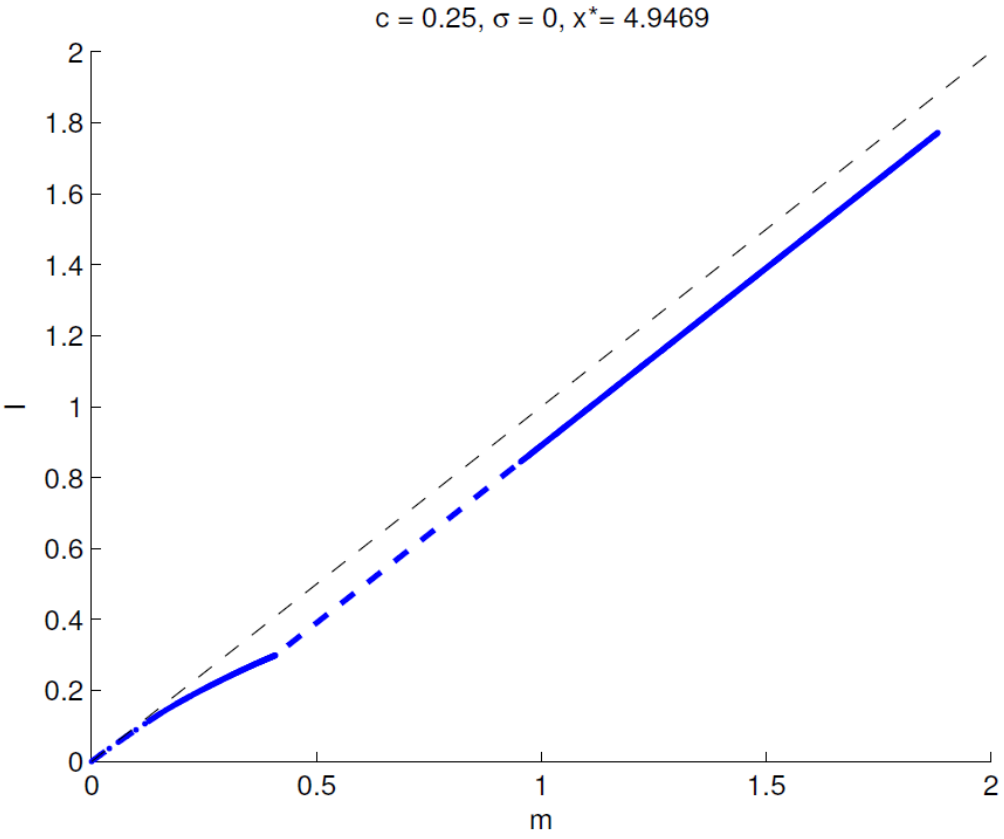




**Figure 3**

**Exponential distribution: Auditing without loading**

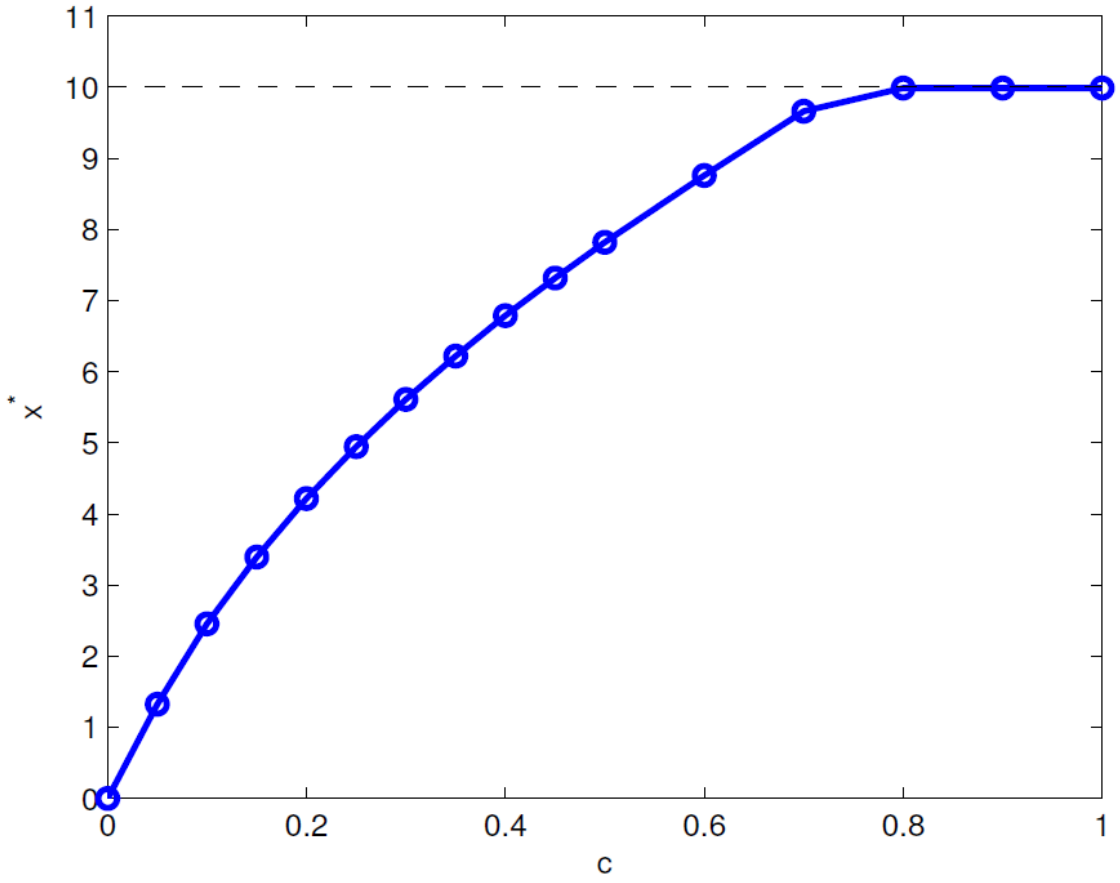



**Figure 4**

**Dependency between threshold $x^*$ and audit cost $c$**

## 5.1 Correlated background risk

Let us first consider the case where the health level affects monetary income through an uninsurable background risk.[30] We assume that illness severity $x$ randomly reduces the monetary wealth by an amount $\varepsilon$. $G(\varepsilon \mid x)$ denotes the c.d.f. of $\varepsilon$, conditionally on $x$ and we assume $G'_x(\varepsilon \mid x) < 0$, where subscript $x$ refers to the partial derivative.[31] Thus, an increase in the illness severity level $x$ shifts the distribution function of $\varepsilon$ in the sense of first-order dominance. Now, the individual's utility is written as $u(R-\varepsilon)+H$, where $R$ denotes the monetary wealth excluding the background risk, and we have

$$V(x, \widetilde{x}) = \overline{u}(R(\widetilde{x}), x) + h_0 - \gamma x[1 - v(m(\widetilde{x}))],$$

still with $R(\widetilde{x}) \equiv w - P + \widehat{I}(\widetilde{x}) - m(\widetilde{x})$, where

$$\overline{u}(R, x) \equiv \int_0^{+\infty} u(R - \varepsilon) dG(\varepsilon \mid x).$$

Thus, the utility of wealth is now written as a state dependent function $\overline{u}(R, x)$, with $\overline{u}'_R > 0, \overline{u}''_{R^2} < 0, \overline{u}'_x < 0$ and $\overline{u}''_{Rx} > 0$. Lemma 2 straightforwardly extends Lemma 1 to this case.

**Lemma 3** *Under correlated background risk, the direct revelation mechanism* $\{m(.), \widehat{I}(.)\}$ *is incentive compatible if and only if*

$$\widehat{I}'(x) = \left[1 - \frac{\gamma x v'(m(x))}{\overline{u}'_R(R(x), x)}\right] m'(x),$$
$$m'(x) \geq 0,$$

*for all* $x \in [0, a]$, *with* $R(x) \equiv w - P + \widehat{I}(x) - m(x)$.

[30] An example is when the individual may lose a part of her business or wage income when her health level deteriorates.

[31] If $\varepsilon$ is continuously distributed, then $G'_\varepsilon(\varepsilon \mid x) > 0$ is the density of $\varepsilon$ conditionally on $x$.

Thus, the necessary and sufficient conditions for incentive compatibility are almost unchanged: we just have to replace $u(R)$ with the state-dependent utility $\overline{u}(R,x)$. Proposition 1, 2 and 3 can be adapted to the case where the individual incurs a correlated background risk, with unchanged conclusion, i.e., the fact that the optimal indemnity schedule does not include a deductible and that bunching at the top may be optimal. Corollary 2 is still valid, but not Corollary 1. In other words, bunching may now be optimal when $x$ is uniformly distributed. Indeed, simulations show that correlated background risk reinforces the likelihood of bunching. We simulate the optimal contract under the assumption $\varepsilon \equiv kx/(a-x) = \varepsilon(x)$ and $\overline{u}(R(x),\varepsilon) = u(R(x)-\varepsilon(x))$,where parameter $k$ measures the intensity of the background risk. Figure 5 illustrates a case where $k = 0.01$ with bunching for the optimal contract.[32]

**Figure 5**

## 5.2 Non-separable utility

We now turn to the case where $U(R,H)$ may be non-separable between $R$ and $H$.[33] It is assumed that $U(R,H)$ is increasing with respect to $R$ and $H$ and concave. We thus have $U'_R > 0, U'_H > 0, U''_{R^2} < 0, U''_{H^2} < 0$ and $U''_{R^2}U''_{H^2} - U''^2_{RH} > 0$. We also assume $U''_{HR} > 0$,[34] and we denote $\Phi(R,H) \equiv U'_H/U'_R$ the marginal rate of substitution between monetary wealth and health, with

$$\Phi'_R = \frac{U''_{HR}U'_R - U'_H U''_{R^2}}{U'^2_R} > 0,$$
$$\Phi'_H = \frac{U''_{H^2}U'_R - U'_H U''_{HR}}{U'^2_R} < 0.$$

[32] In Figure 5-top, indifference curves for $x = 7$ and 9 almost coincide. Figure 5-bottom shows that $\overline{m}$ decreases when $k$ increases, with a decrease in the upper limit of the insurance indemnity $I(\overline{m})$. There is bunching only when $k > 0$ since Figure 5 corresponds to the case of uniform distribution.

[33] Henceforth, we assume there is no background risk.

[34] $U''_{HR} > 0$ is assumed for the sake of simplicity. Lemma 3 is valid under more general conditions that are compatible with $U''_{HR} \leq 0$.

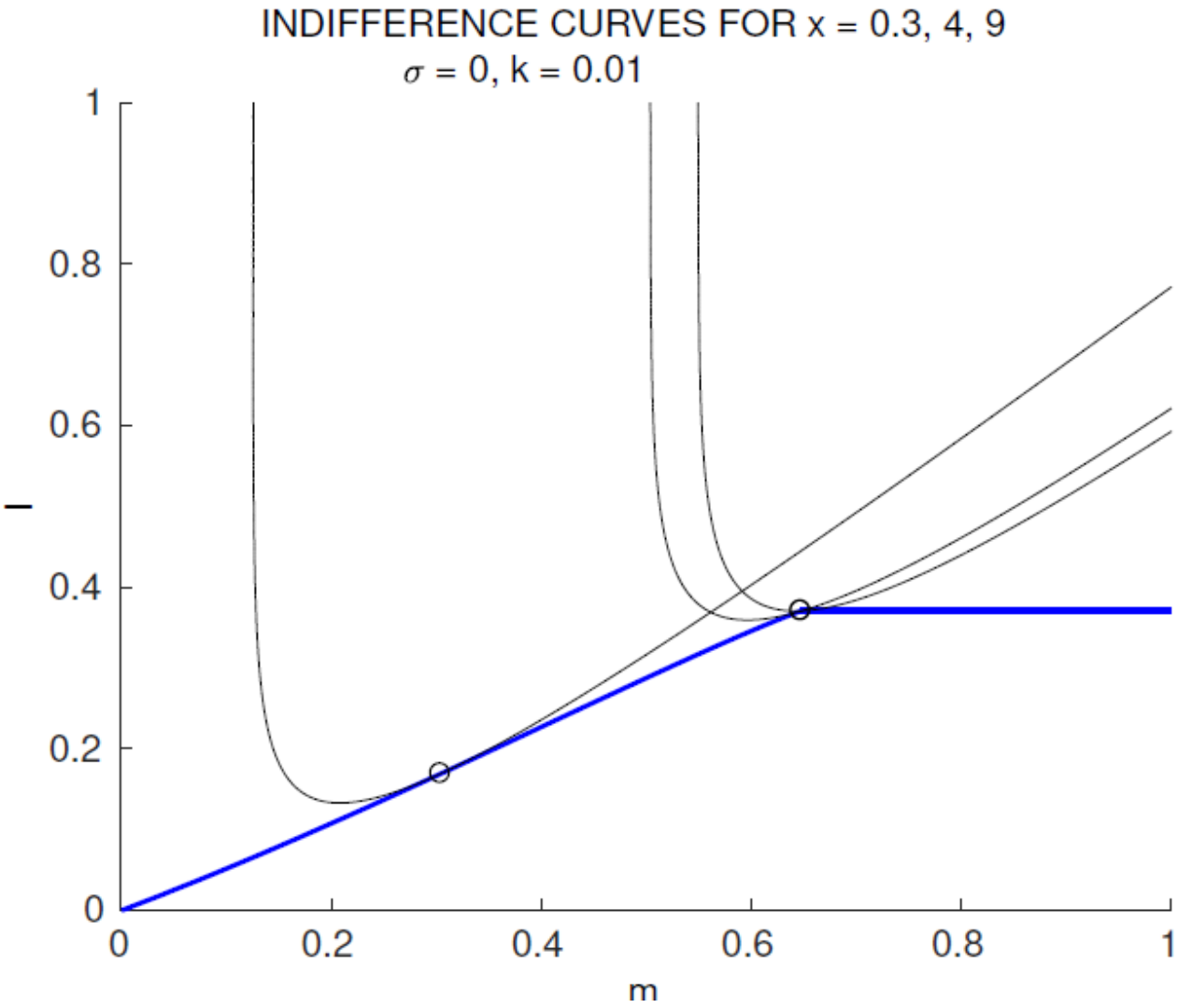



**Figure 5**

**Uniform distribution**

**Case where the background risk creates bunching**

Thus, the individual is more willing to pay for a marginal improvement in her health level when her income is higher and when her health level is lower. We now have

$$V(x,\widetilde{x}) = U\left(w - P + \widehat{I}(\widetilde{x}) - m(\widetilde{x})), h_0 - \gamma x[1 - v(m(\widetilde{x}))]\right).$$

Lemma 4 is a direct extension of Lemma 1 to the case of a non-separable utility function, with a similar interpretation.

**Lemma 4** *Under non-separable utility, the direct revelation mechanism* $\{m(.), \widehat{I}(.)\}$ *is incentive compatible if and only if*

$$\widehat{I}'(x) = \left[1 - \gamma x v'(m(x))\Phi\left(R(x), H(x)\right)\right] m'(x), \tag{17}$$

$$m'(x) \geq 0, \tag{18}$$

*for all* $x \in [0,a]$*, where* $R(x) \equiv w - P + \widehat{I}(x) - m(x)$ *and* $H(x) \equiv h_0 - \gamma x[1 - v(m(x))$.

Now, the optimal incentive compatible mechanism maximizes

$$\int_0^a \left\{U\left(w - P + \widehat{I}(x) - m(x)), h_0 - \gamma x(1 - v(m(x)))\right)\right\} f(x)dx$$

with respect to $\widehat{I}(.), m(.), h(.)$ and $P$, subject to $\widehat{I}(0) = 0$, and (2),(7)-(9), and

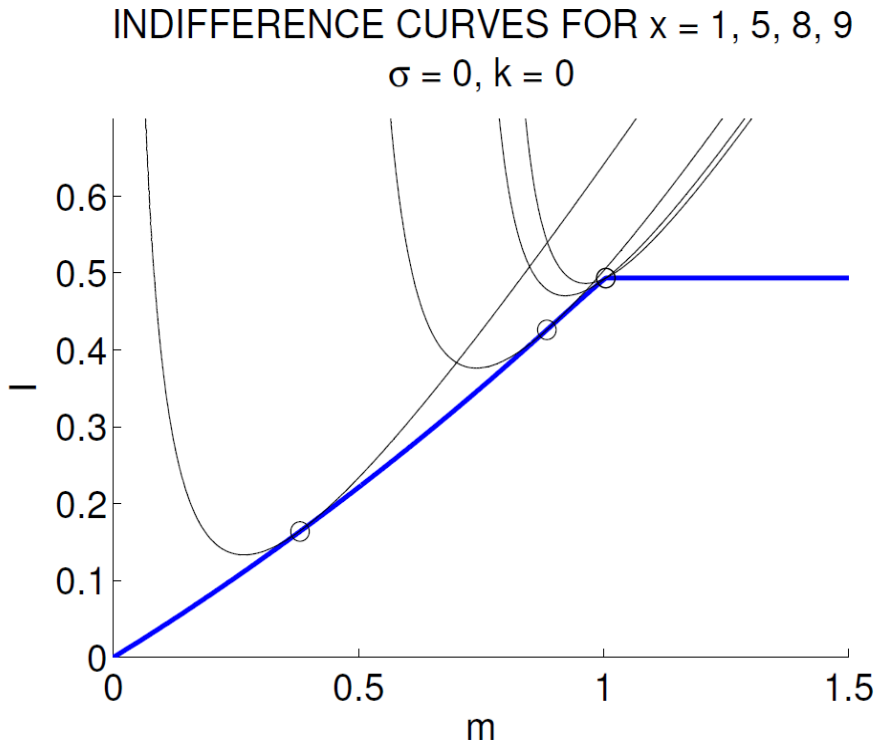
$$\widehat{I}'(x) = \left[1 - \gamma x v'(m(x))\Phi\left(R(x), H(x)\right)\right] h(x). \tag{19}$$

We have simulated the non-separable utility case with $U(R,H) = (b_0 \frac{R^{1-\alpha}}{1-\alpha} + b_1)H^{\beta}$.[35] The optimal indemnity schedule remains qualitatively similar to the characterization provided in Section 2. Figure 6-top illustrates the case of an exponential distribution with bunching.[36]

**Figure 6**

[35] Thus, utility is CRRA w.r.t. wealth. Parameters are $\alpha = 2, \beta = 0.5, b_0 = 0.01$ and $b = 1$.

[36] Figure 6-bottom adds a background risk and a loading factor, and it illustrates the optimality of a deductible, as shown in Section 5.3.



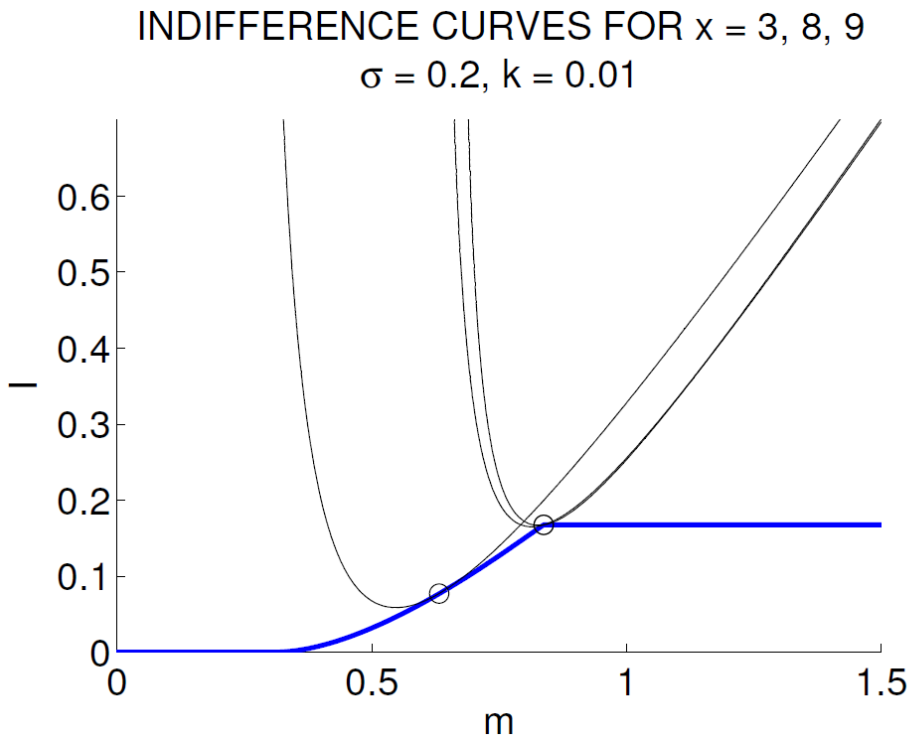




**Figure 6**

**Non-separable utility – Exponential distribution**

## 5.3 Insurance loading

In practice, insurance pricing includes a loading that reflects various underwriting costs, including commissions to agents and brokers, operating expenses, loss adjustment expenses and capital cost. Let us assume that the premium is loaded at rate $\sigma$, which gives

$$P = (1+\sigma)\int_0^a \widehat{I}(x)f(x)dx, \tag{20}$$

instead of (2). As initially established by Arrow (1971), the optimal contract contains a straight deductible when there is a positive constant loading factor. Propositions 6 and 7 extend this characterization to the case of ex post moral hazard.

**Proposition 6** *Under constant positive loading $\sigma$ and with the same assumptions as Proposition 2, the optimal indemnity schedule without auditing includes a deductible $D > 0$ and an upper limit $I(\overline{m})$, that is*

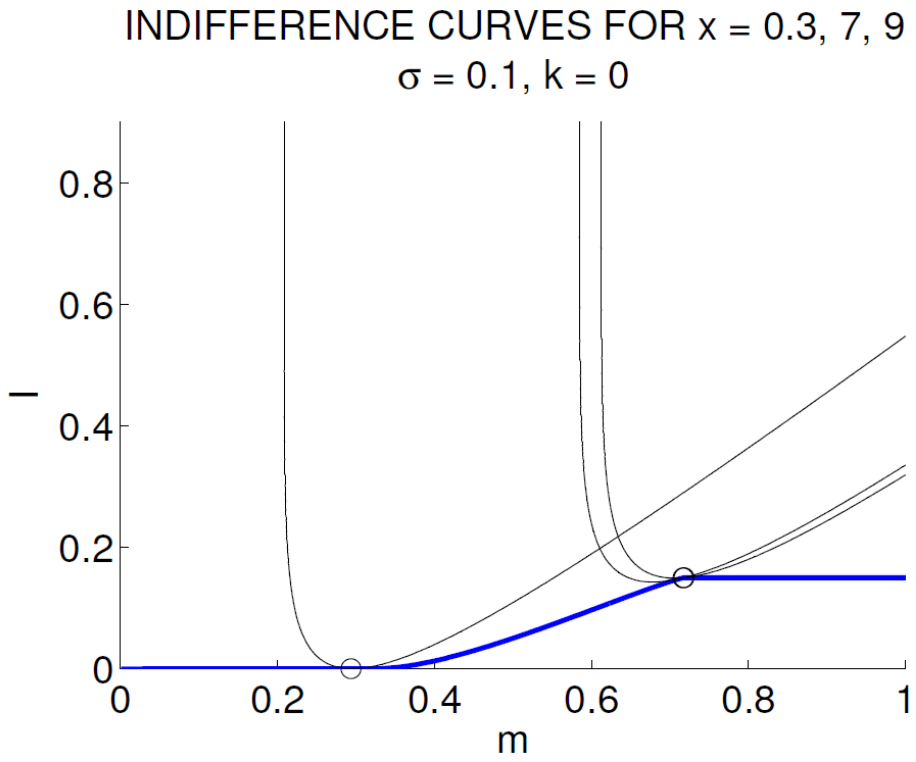
$$\begin{aligned}
&I(m) = 0 \quad \textit{if} \quad m \leq D,\\
&I'(D) \in [0,1),\\
&I'(m) \in (0,1) \ \textit{if} \ \ m \in [D,\overline{m}),\\
&I(m) = I(\overline{m}) \quad \textit{if} \quad m \geq \overline{m},\\
&I'(\overline{m}) = 0 \ \textit{if}\, \overline{x} = a, I'(\overline{m}) > 0 \ \textit{if}\, \overline{x} < a.
\end{aligned}$$

**Corollary 3** *Under the same assumptions as Corollary 1, we have $\overline{x} = a$, i.e., there is no bunching.*

**Corollary 4** *Under the same assumptions as Corollary 2, we have $\overline{x} < a$, i.e., there is bunching.*

Figure 7 illustrates Corollary 4 in the case of an exponential distribution. Loading shifts the indemnity schedule rightward and creates a deductible ($D \simeq 0.3202 = m(0.41)$ when $\sigma = 0.1$), in addition to bunching at the top.

**Figure 7**



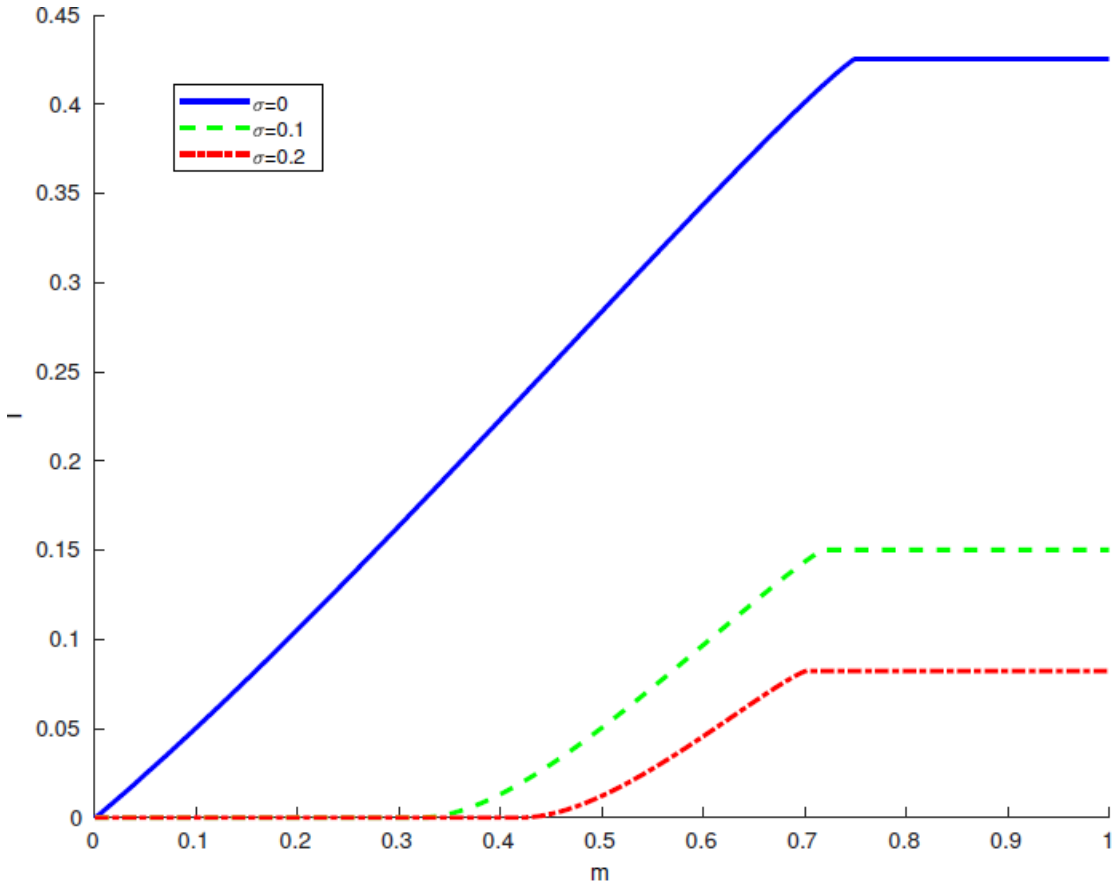




**Figure 7**

**Exponential distribution - A deductible is optimal under loading**

# 6 Public policy objectives

## 6.1 Income redistribution

So far, we have assumed that all patients have identical wealth $w$. From a normative standpoint, this refers either to an unrealistic world where social inequality does not exist, or to a case where policy instruments allow the government to design health insurance and income redistribution independently. More realistically, the literature on optimal taxation highlights that income redistribution is limited by incentive compatibility and informational constraints. This leads to the persistence of wealth inequalities that may affect how health insurance should be designed.

Analyzing a fully-fledged integrated public policy, with interaction between income redistribution and health care insurance coverage, would go far beyond the objectives of this paper. We may nevertheless draw upon our model to sketch out how income inequalities may affect our conclusions. A possible approach consists in assuming that there are several groups of policyholders indexed by $i = 1, ..., n$, with initial wealth $w_i$ in group $i$ and identical health risk exposures and risk preference. In other words, all groups are identical, except as regards initial wealth. Let $\theta_i \geq 0$ be the fraction of type $i$ individuals in the population, with $\sum_{i=1}^{n} \theta_i = 1$. Initial wealth $w_i$ is assumed to be perfectly observable and health insurance is provided through group-specific insurance contracts, with indemnity schedule $I_i(m)$, premium $P_i$ and state dependent indemnity $\widehat{I}_i(x)$ and health expenses $m_i(x)$ for group $i$. The government can cross-subsidize insurance contracts through net subsidies $s_i$ per type $i$ policyholder, $-s_i$ being a tax if $s_i < 0$, while meeting a budget constraint

$$\sum_{i=1}^{n} \theta_i s_i = B,$$

where $B$ denotes the amount of financial resources allocated by the government to health insurance, possibly with $B = 0$. The insurer's break-even condition for group $i$

is written as

$$\widetilde{P}_i \geq \int_0^a \widehat{I}_i(x) f(x) dx.$$

where $\widetilde{P}_i = P_i + s_i$, and the net income of type $i$ policyholders with health expenses $m$ is written as

$$\begin{aligned} R_i &= w_i - P_i - m + I_i(m) \\ &= \widetilde{w}_i - \widetilde{P}_i - m + I_i(m), \end{aligned}$$

where $\widetilde{w}_i = w_i + s_i$. Hence, for given net subsidies $s_i$, everything happens as if type $i$ policyholders purchase optimal insurance at actuarial price $\widetilde{P}_i$ with adjusted net wealth $\widetilde{w}_i = w_i + s_i$. The limits to income redistribution are not analyzed here and they lead to heterogeneous adjusted net wealth levels $\widetilde{w}_i$ among the groups $i = 1, ..., n$. Thus, when income redistribution is endogenously defined by net subsidies $s_i$, analyzing the effects of incomplete income redistribution on health insurance boils down to considering the relationship between adjusted net wealth and optimal health coverage.

The effects of wealth on optimal health insurance under ex post moral hazard go through several mechanisms that are difficult to disentangle. Firstly, the degree of risk aversion may be affected by wealth, and in particular under the DARA assumption, higher wealth involves a lower absolute risk aversion and thus, a larger propensity to purchase insurance. Secondly, larger wealth induces larger marginal willingness to pay for health care, which is reflected in higher health expenses for any health state, and thus in higher risk exposure. This also justifies paying a higher insurance premium, without an obvious predictable conclusion on the shape of the indemnity schedule. Thirdly, in an ex post moral setting, decreasing the insurance coverage acts as a self-discipline device to moderate health expenses, which may be ultimately beneficial to wealthy policyholders since they have a high propensity to spend money for health care. These effects interact in a complex way, and we have to rely on simulations to draw some conclusions. Figures 8 and 9 illustrate the optimal health expenses functions $m(x)$ and the optimal indemnity schedule $I(m)$ for various levels of initial

wealth.[37] Figure 8 shows that increasing the initial wealth shifts the graph of $m(x)$ upwards: for a given health state $x$, the larger the initial net wealth $w$, the larger the expenditure for health care $m(x)$. This is not astonishing since an increase in wealth goes along with an increase in the marginal willingness to pay for health care. As shown in Figure 9,[38] the increase in wealth has very little impact on the indemnity schedule: increasing initial wealth slightly shifts the graph of function $I(m)$ upwards, but this impact is barely noticeable, particularly for low expense levels.[39]

**Figures 8 and 9**

Although these simulations should not be overinterpreted, they suggest that the redistributive objective of public policy should be focused on direct transfers to individuals (through the net subsidy $s_i$ to each group $i$) rather than by differentiating the reimbursement of medical expenses. In other words, this leads to the speculation that combining a unique reimbursement schedule with differentiated direct transfers to individuals may adequately approximate the optimal strategy of a government with redistributive objective. Assessing the validity of this conjecture would require a detailed analysis, where the optimal design of health insurance and income redistribution would be simultaneously considered.

## 6.2 Agency relationship between physician and patient

We have reduced attention to a simple setting where the patient is perfectly well informed about her health state and where health care providers act as "perfect agents" with the only objective of maximizing the patient's utility. In practice, the asymmetry of information between patient and insurer may be combined with a physician-patient

[37] Figures 8 and 9 correspond to the assumptions made in Section 3.3 without loading or background risk, when the distribution of $x$ is exponential.

[38] The indifference curves are drawn for $x = 2$ in Figure 9.

[39] See Picard (2016) for a case with linear coinsurance where this effect of wealth on the coinsurance rate vanishes completely.

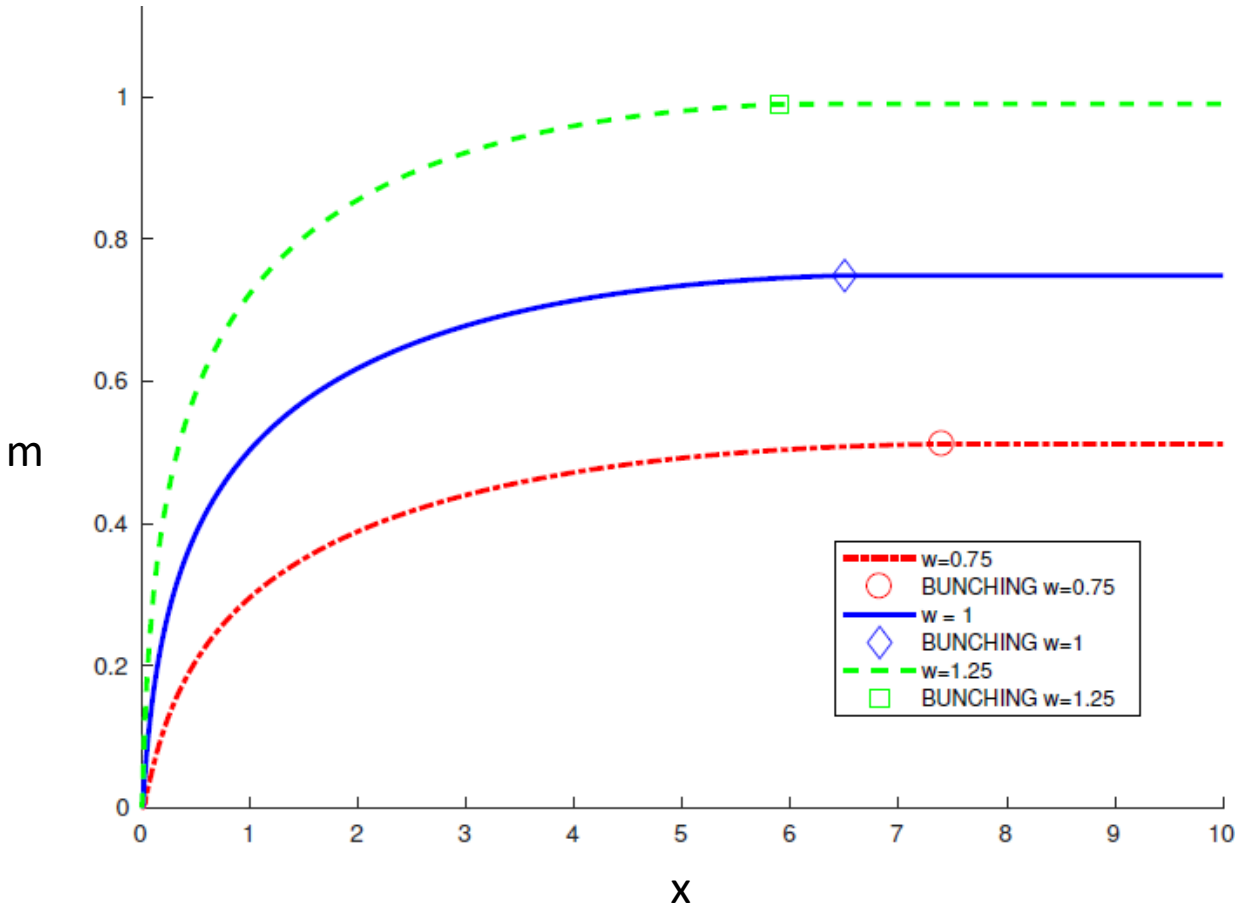



**Figure 8**

**Health expense profile for various levels of wealth**

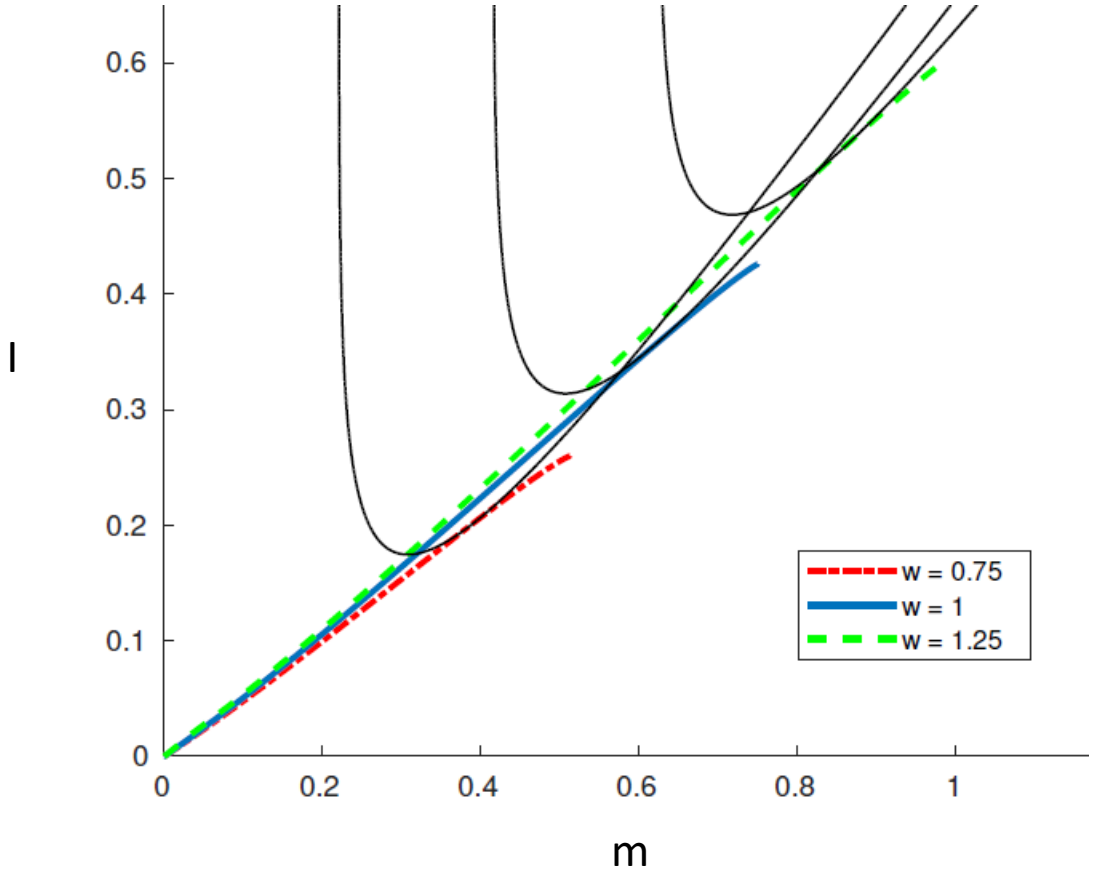



**Figure 9**

**Indemnity schedule and indifference curves**

**for various levels of wealth**

agency relationship, where the patient has incomplete information about her health state and the physician may over-prescribe and over-treat either to increase income, or to avoid the risk of being accused of negligence. This issue raises questions about the relationship between the design of insurance schemes and the functioning of health care market. It is of utmost importance for social security and private insurers, and it may lead insurers to monitor the behavior of health care providers through various forms of health care management plan, providers affiliated networks or vertical integration.

Let us sketch an extension of our analysis that illustrates the effects of the physician-patient agency relationship. Assume that only a part $q(m)$ of total medical expenses $m$ is really useful, with $0 < q(m) < m$ and $q'(m) > 0$. The complement $m - q(m)$ corresponds to efficiency losses induced by imperfect monitoring of health care providers. Assume $q''(m) > 0$. Hence $q(m)/m$ is increasing, which expresses the idea that, in proportion to total medical costs, excessive expenses mainly correspond to an excessive number of minor medical acts.[40] The policyholder's utility is written by substituting $v(q(m))$ to $v(m)$ in the policyholder's expected utility. All the other assumptions of section 3.3 are unchanged, and in particular we assume $v(m) = \sqrt{m}/[1+\sqrt{m}]$. Figures 10, 11 and 12 illustrate the consequences of this assumption, by assuming $q(m) = m^\alpha$, with $\alpha = 1, 2, 3$ and $4$.[41]

**Figures 10, 11, 12 and 13**

The locus of function $v(m^\alpha)$ is drawn in Figure 10. It highlights the increase in the efficiency loss when $\alpha > 1$, by comparison with the benchmark case $\alpha = 1$. Figure 10 also illustrates the fact that $v(m^\alpha)$ is no longer concave when $\alpha$ is large enough (equal to 3 or 4). Figures 11 and 12 show that the inefficiency of small medical expenses leads to reduce these expenses and the corresponding insurance coverage (and even to fully cancel them when $x$ is small and $\alpha$ is equal to 3 or 4)[42] and to increase expenses

[40] This is just an assumption made for illustrative purposes.

[41] The relevant values are such $m < 1$, and thus $q(m) = m^\alpha < m$.

[42] This corner solution is induced by the non concavity of $v(m^\alpha)$ when $\alpha = 3$ and 4.

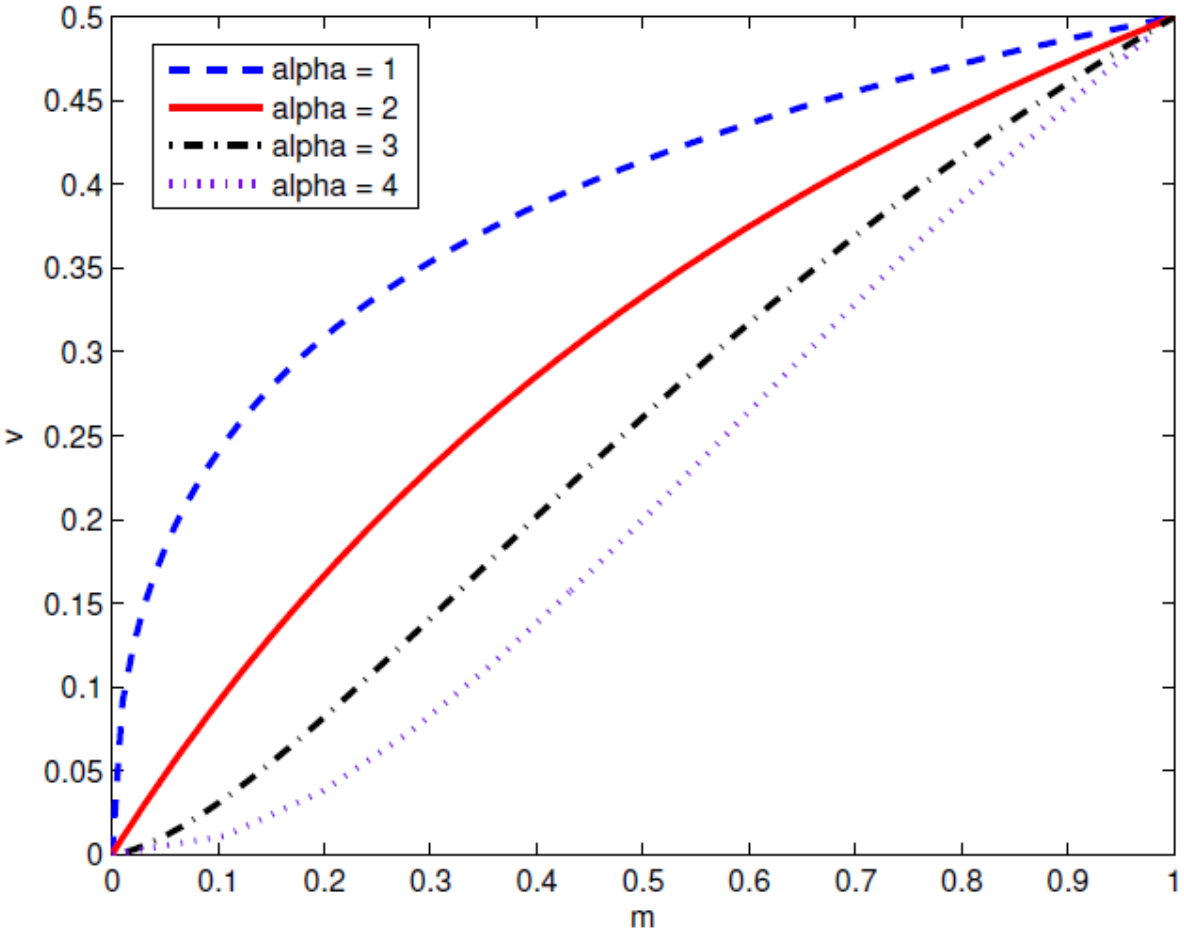



**Figure 10**

**Inefficiency from the physician-patient agency relationship**

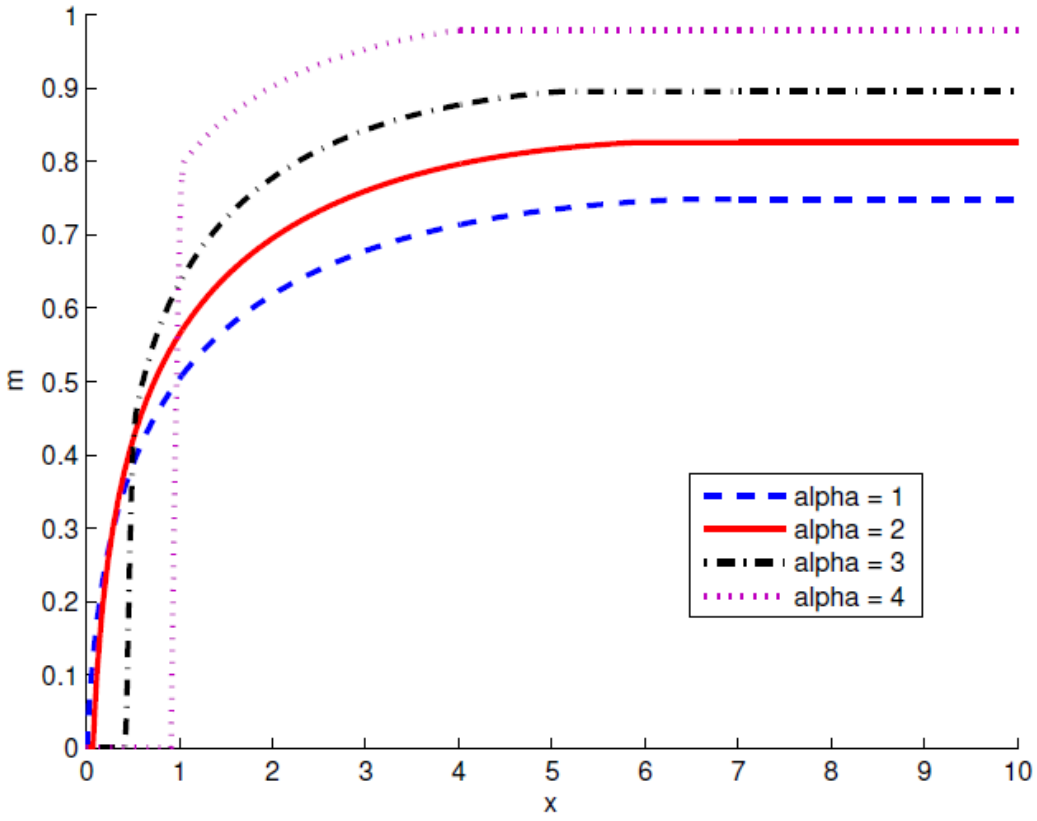



**Figure 11**

**Health expense profile with agency costs**

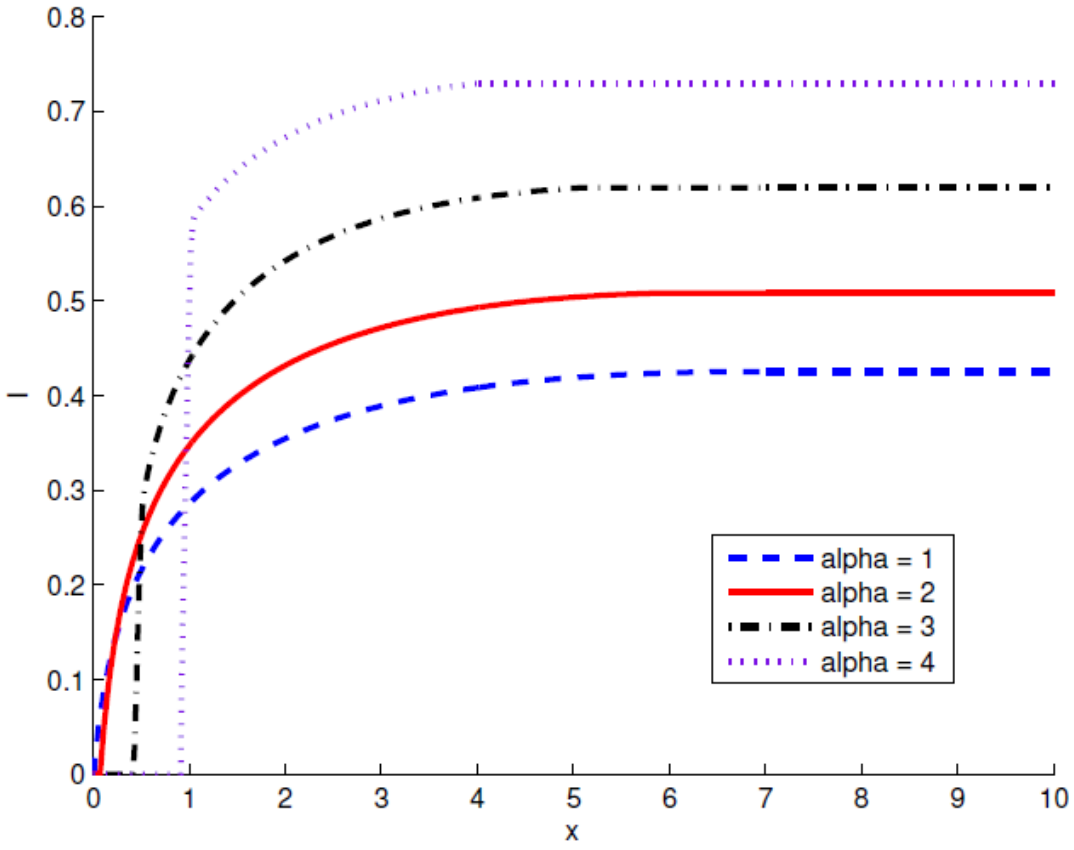



**Figure 12**

**Insurance indemnity profile with agency costs**

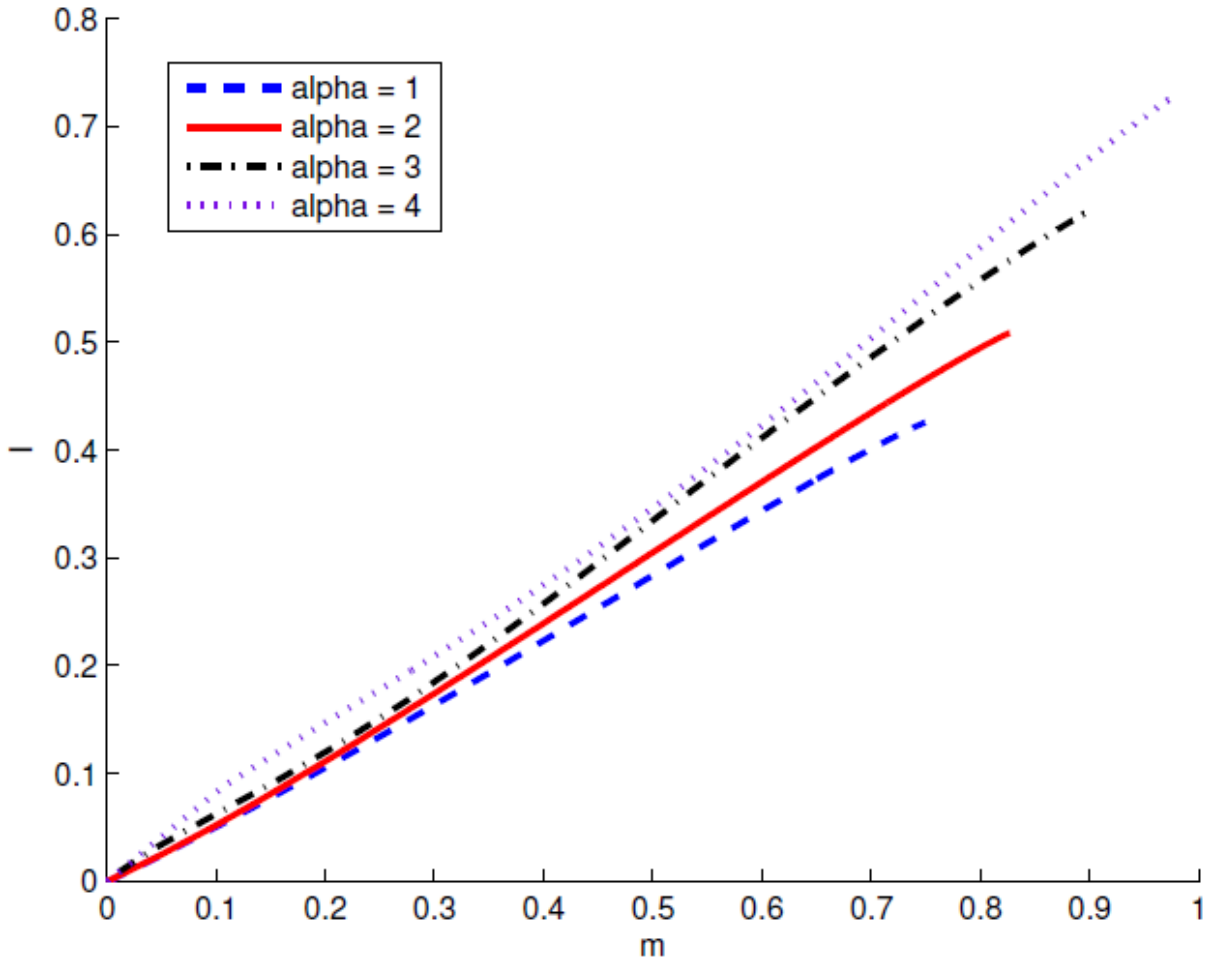



**Figure 13**

**Indemnity schedule with agency costs**

and coverage when $x$ is large. Figure 13 shows that changes in parameter $\alpha$ induce relatively small shifts in the indemnity schedule $I(m)$.

# 7 Conclusion

Using demand management to mitigate the consequences of ex post moral hazard in medical insurance goes through an adequate definition of the indemnity schedule. Under reasonable assumptions, it has been shown that the optimal solution mixes partial coverage at the margin and an upper limit to coverage under the form of bunching: the most acute types of illness severity lead to the same expenses and to the same insurance indemnity. Our second main result is about the optimality of a deductible. A deductible may be optimal only if the insurer charges loaded premiums. In other words, deductibles should not be part of the solution to the incentive-risk sharing trade-off in itself. They are the consequence of transaction costs reflected in an insurance loading factor, and they reflect the level of these costs. This is an important difference between ex post and ex ante moral hazard. Finally, we have immersed our ex post moral hazard problem in a costly state verification setting where the insurer can monitor the health expenses through auditing. We have shown that there should be coinsurance at the margin, and possibly an upper limit to coverage, when the sickness severity is lower than a threshold under which there is no audit. When the sickness severity crosses this limit, then it is optimal to audit the health state, with an upward jump in care expenses. In this regime, there is full insurance at the margin, which corresponds to an out-of-pocket maximum.

Overall, this analysis reveals a contrasting picture of the way health expenses should be reimbursed by insurers. On the one hand, there are limits to coverage for low expenses under the combination of coinsurance, upper limit and deductible. On the other hand, the largest expenses should be more generously covered, with upper limits to out-of-pocket expenses. This complexity reflects what we frequently observe in the

real world when all these ingredients are mixed, with more complete coverage and out-of-pocket maximum, for large easily monitorable categories such as surgery or other forms of inpatient care, and coinsurance or upper limits that aim to contain health spending for minor illnesses.[43]

# Appendix 1

**Proof of Lemma 1**

**Step 1**: *There exists an optimal revelation mechanism.*

Let us change variables by denoting $A(x) = u(w - P + \widehat{I}(x) - m(x))$ and $B(x) = v(m(x))$. The incentive compatibility constraints and the insurer's break-even constraint are respectively rewritten as

$$A(x) + \gamma x B(x) \geq A(\widetilde{x}) + \gamma x B(\widetilde{x}) \text{ for all } x, \widetilde{x}, \tag{21}$$

$$w \geq \int_0^a [u^{-1}(A(x)) + v^{-1}(B(x)]f(x)dx, \tag{22}$$

Furthermore, $\widehat{I}(0) = m(0) = 0$ gives $A(0) = u(w - P)$ and $B(0) = 0$. Let $\mathcal{S}$ be the subset of functions $A(.), B(.)$ that belong to the Banach space $\mathcal{L}^\infty([0,1], \mathbb{R} \times [0,1])$ with the sup norm topology $\| \, . \, \|_\infty$ and that satisfy (21),(22) and $B(0) = 0$. Hence, $\mathcal{S}$ is closed and convex, and furthermore $\| \, (A(.), B(.)) \, \|_\infty \leq u(w)$ for all $(A(.), B(.)) \in \mathcal{S}$. Let

$$J = \int_0^a \{A(x) + h_0 - \gamma x[1 - B(x)]\}f(x)dx.$$

$J$ is a linear (and thus weakly concave) function of $A(.), B(.)$. Hence, it reaches a maximum in $\mathcal{S}$, which proves the existence of an optimal incentive compatible mechanism, with $P = w - u^{-1}(A(0))$.

**Step 2:** *For any incentive compatible mechanism, $m(x)$ and $\widehat{I}(x)$ are non-decreasing.*

Incentive compatibility implies

$$u(w - P - m(x) + \widehat{I}(x)) - u(w - P - m(\widetilde{x}) + \widehat{I}(\widetilde{x})) \geq \gamma x[v(m(\widetilde{x})) - v(m(x))],$$

[43] For the sake of illustration, see for instance Kaiser Family Foundation (2009) for France, Germany and Switzerland, and www.healthcare.gov for the ObamaCare Marketplace in the US.

and, reversing the roles of $x$ and $\widetilde{x}$,

$$u(w-P-m(x)+\widehat{I}(x))-u(w-P-m(\widetilde{x})+\widehat{I}(\widetilde{x}))\leq\gamma\widetilde{x}[v(m(\widetilde{x}))-v(m(x))].$$

We deduce $(\widetilde{x}-x)[v(m(\widetilde{x}))-v(m(x))]\geq 0$ for all $x,\widetilde{x}$, which implies that $m(.)$ is non-decreasing. Since $I(.)$ is non-decreasing, $\widehat{I}(.)\equiv I(m(.))$ is also non-decreasing.

**Step 3**: *For any optimal revelation mechanism,* $m(.)$ and $\widehat{I}(.)$ *are continuous.*

Let $\{m_0(.),\widehat{I}_0(.)\}$ be an optimal incentive compatible revelation mechanism and suppose that $m_0(.)$ is rightward discontinuous[44] at $x_*\in(0,a)$, with $m_0(x)\to m_0(x_*)+\Delta_m$ and $\widehat{I}_0(x)\to\widehat{I}_0(x_*)+\Delta_I$, when $x\to x_*, x>x_*$, with $\Delta_m>0$ and $\Delta_I\geq 0$. Incentive compatibility implies that a type $x_*$ individual is indifferent between $m_0(x_*),\widehat{I}_0(x_*)$ and $m_0(x_*)+\Delta_m,\widehat{I}_0(x_*)+\Delta_I$. If $\Delta_I=0$, since $I(m)$ is non-decreasing, it remains constant when $m\in[m_0(x_*),m_0(x_*)+\Delta_m]$. Using the concavity of $m\to u(w-P-m+\widehat{I}_0(x_*))+\gamma x_* v(m)$ then shows that the type $x_*$ individual reaches a higher expected utility by choosing $m\in(m_0(x_*),m_0(x_*)+\Delta_m)$ than by choosing $m_0(x_*)$, hence a contradiction. Thus, we have $\Delta_I>0$.

Since $\widehat{I}_0(x)$ is piecewise continuous, there exists $\theta>0$ such that $\widehat{I}_0(x)-\widehat{I}_0(x_*)\geq\Delta_I/2$ for all $x\in(x_*,x_*+\theta)$. Consider another revelation mechanism $\{m_1(.),\widehat{I}_1(.)\}$ defined by:

$(i)$ If $x\in(x_*,x_*+\theta)$, let $m_1(x)=m_1^*$ and $\widehat{I}_1(x)=I_1^*$ close to $m_0(x_*)$ and $\widehat{I}_0(x_*)$, respectively, with $\widehat{I}_0(x)-I_1^*\geq\Delta_I/4$, and such that

$$u(w-P-m_1^*+I_1^*)+\gamma x v(m_1^*)\geq u(w-P-m_0(x)+\widehat{I}_0(x))+\gamma x v(m_0(x)),$$

for all $x\in(x_*,x_*+\theta)$, and

$$u(w-P-m_1^*+I_1^*)+\gamma x v(m_1^*)<u(w-P-m_0(x)+\widehat{I}_0(x))+\gamma x v(m_0(x)),$$

if $x\leq x_*$,

$(ii)$ If $x\notin(x_*,x_*+\theta)$, then $m_1(x)\equiv m_0(x)$ and $\widehat{I}_1(x)\equiv\widehat{I}_0(x)$.

[44] A similar proof applies to the case of leftward discontinuity.

Let $\widetilde{x}_1(x)$ be an optimal report of a type $x$ policyholder in $\{m_1(.), \widehat{I}_1(.)\}$, with $\widetilde{x}_1(x) = x$ for all $x \in [0, x_* + \theta)$, and let $\{m_2(.), \widehat{I}_2(.)\}$ be the incentive compatible revelation mechanism defined by $m_2(x) \equiv m_1(\widetilde{x}_1(x)), \widehat{I}_2(x) \equiv \widehat{I}_1(\widetilde{x}_1(x))$. For $P$ unchanged, the policyholder's expected utility is higher for $\{m_2(.), \widehat{I}_2(.)\}$ than for $\{m_0(.), \widehat{I}_0(.)\}$. Furthermore, $\widehat{I}_2(x) = \widehat{I}_0(x)$ if $x < x^*$, $\widehat{I}_2(x) = I_1^* < \widehat{I}_0(x) - \Delta_I/4$ if $x_* \leq x < x_* + \theta$ and $\widehat{I}_2(x) \leq \widehat{I}_0(x)$ if $x \geq x_* + \theta$. Hence, $\{m_2(.), \widehat{I}_2(.)\}$ is feasible with $P$ unchanged, which contradicts the optimality of $\{m_0(.), \widehat{I}_0(.)\}$.

**Step 4**: *(4) and (5) are necessary and sufficient conditions for a continuous revelation mechanism to be incentive compatible.*

Local first-order and second-order incentive compatibility conditions for type $x$ are written respectively as

$$\frac{\partial V(x, \widetilde{x})}{\partial \widetilde{x}} \mid_{\widetilde{x}=x} = 0, \tag{23}$$

$$\frac{\partial^2 V(x, \widetilde{x})}{\partial \widetilde{x}^2} \mid_{\widetilde{x}=x} \leq 0, \tag{24}$$

at any point of differentiability. (23) and (24) are necessary conditions for incentive compatibility. We have

$$\frac{\partial V(x, \widetilde{x})}{\partial \widetilde{x}} = u'(R(\widetilde{x}))[\widehat{I}'(\widetilde{x}) - m'(\widetilde{x})] + \gamma x v'(m(\widetilde{x}))m'(\widetilde{x}),$$

and thus (23) yields (4).

Since (4) should hold for all $x \in [0, a]$, a simple calculation yields

$$\frac{\partial^2 V(x, \widetilde{x})}{\partial \widetilde{x}^2} \mid_{\widetilde{x}=x} = -\gamma v'(m(x))m'(x),$$

and thus (24) gives (5).

Conversely, assume (4) and (5) hold. (4) gives

$$\frac{\partial V(x, \widetilde{x})}{\partial \widetilde{x}} = \gamma(x - \widetilde{x})v'(m(\widetilde{x}))m'(\widetilde{x}).$$

Using (5) then shows that $\partial V(x, \widetilde{x})/\partial \widetilde{x} \leq 0$ if $\widetilde{x} > x$ and $\partial V(x, \widetilde{x})/\partial \widetilde{x} \geq 0$ if $\widetilde{x} < x$, which implies incentive compatibility.

**Proof of Proposition 1**

Let $\mu_1(x)$ and $\mu_2(x)$ be costate variables for $\widehat{I}(x)$ and $m(x)$ respectively, and let $\lambda$ and $\delta(x)$ be Lagrange multipliers respectively for (2) and (9). The Hamiltonian is written as

$$\mathcal{H} = [u(R(x)) + \gamma x v(m(x))]f(x) + \mu_1(x)h(x)\left[1 - \frac{\gamma x v'(m(x))}{u'(R(x))}\right] + \mu_2(x)h(x) - \lambda\widehat{I}(x)f(x) + \delta(x)\widehat{I}(x).$$

The optimality conditions are

$$\varphi(x) \equiv \mu_1(x)\left[1 - \frac{\gamma x v'(m(x))}{u'(R(x))}\right] + \mu_2(x) \leq 0, = 0 \text{ if } h(x) > 0, \tag{25}$$

$$\mu_1'(x) = [\lambda - u'(R(x))]f(x) - \mu_1(x)h(x)\gamma x\frac{v'(m(x))u''(R(x))}{u'(R(x))^2} - \delta(x), \tag{26}$$

$$\mu_2'(x) = [u'(R(x)) - \gamma x v'(m(x))]f(x) + \mu_1(x)h(x)\gamma x\left[\frac{v''(m(x))u'(R(x)) + v'(m(x))u''(R(x))}{u'(R(x))^2}\right], \tag{27}$$

$$\mu_1(a) = \mu_2(a) = 0, \tag{28}$$

$$\lambda - \int_0^a \left[u'(R(x))f(x) + \mu_1(x)h(x)\gamma x\frac{v'(m(x))u''(R(x))}{u'(R(x))^2}\right]dx = 0, \tag{29}$$

with $\delta(x) \geq 0$ and $\delta(x) = 0$ if $\widehat{I}(x) > 0$. A tedious but straightforward calculation using (26) and (27) leads to

$$\varphi'(x) = [\lambda f(x) - \delta(x)]\left[1 - \frac{\gamma x v'(m(x))}{u'(R(x))}\right] - \gamma\mu_1(x)\frac{v'(m(x))}{u'(R(x))}. \tag{30}$$

We also have $R'(x) = \widehat{I}'(x) - m'(x) = -\gamma x h(x)v'(m(x))/u'(R(x)) \leq 0$. Thus, $R(x)$ is non-increasing, and it is decreasing when $h(x) > 0$. The remaining part of the proof is in five steps.

**Step 1**: $m(x) > 0$ *for all* $x > 0$.

Since $m(0) = 0$ and $m(x)$ is non-decreasing, there exists $\underline{x} \in [0, a]$ such that $m(x) > 0$ if and only if $x > \underline{x}$. Suppose $\underline{x} > 0$, which implies $h(x) = 0$ over $[0, \underline{x}]$.

Using $\widehat{I}(0) = 0$ and (6) gives $\widehat{I}(x) = 0$ for all $x \in [0, \underline{x}]$. Let

$$\widehat{m}(x) \equiv \arg\max_{\widetilde{m} \geq 0} \{u(w - P - \widetilde{m}) + \gamma x v(\widetilde{m})\}, \tag{31}$$

with $\widehat{m}(x) > 0$ for all $x > 0$. Define $m_0(x) = \widehat{m}(x), I_0(x) = 0$ if $x \leq \underline{x}$ and $m_0(x) = m(x), I_0(x) = \widehat{I}(x)$ if $x > \underline{x}$, and

$$x_0(x) \in \arg\max_{\widetilde{x} \in [0,a]} \{u(w - P - m_0(\widetilde{x}) + I_0(\widetilde{x})) + x v(m_0(\widetilde{x}))\}.$$

The revelation mechanism $m_1(.), \widehat{I}_1(.)$ defined by $m_1(x) \equiv m_0(x_0(x))$ and $\widehat{I}_1(x) \equiv I_0(x_0(x))$ is incentive compatible and it dominates the supposed optimal mechanism $m(.), \widehat{I}(.)$ - i.e., it provides a higher expected utility to the policyholder and its expected profit is non-negative for $P$ unchanged -, hence a contradiction. Thus, $\underline{x} = 0$.

**Step 2:** *$\mu_1(x)$ is continuous in $[0, a]$ with $\mu_1(x) = 0$ if $\widehat{I}(x) = 0$.*

Let $x_0 \in (0, a)$ be a junction point such that $\widehat{I}(x) = 0$ if $x \in (x_0 - \varepsilon, x_0]$ and $\widehat{I}(x) > 0$ if $x \in (x_0, x_0 + \varepsilon)$, with $0 < \varepsilon < x_0$.[45]

Using the same argument as in Step 1 shows that $h(x) > 0$ in $(x_0 - \varepsilon, x_0)$. Let $x \in (x_0 - \varepsilon, x_0)$. Using $h(x) > 0, \widehat{I}'(x) = 0$ and (6) gives $u'(R(x)) = \gamma x v'(m(x))$. Then, $\varphi(x) = 0$ gives $\mu_2(x) = 0$ and thus $\mu_2'(x) = 0$ for all $x \in (x_0 - \varepsilon, x_0]$. (30) implies $\mu_1(x) = 0$ for all $x \in (x_0 - \varepsilon, x_0)$, and this is true, more generally, for all $x \in [0, a]$ such that $\widehat{I}(x) = 0$.

Let $x \in (x_0, x_0 + \varepsilon)$. $\widehat{I}(x)$ is locally increasing over $(x_0, x_0 + \varepsilon)$ and thus $\widehat{I}'(x) > 0$ and $h(x) > 0$ (at least for $\varepsilon$ small enough). Thus, we have $\delta(x) = \varphi(x) = \varphi'(x) = 0$ for all $x \in (x_0, x_0 + \varepsilon)$. Since $R(x)$ and $m(x)$ are continuous functions and $u'(R(x_0)) = \gamma x_0 v'(m(x_0))$, we have $u'(R(x)) - \gamma x v'(m(x)) \to 0$ when $x \searrow x_0$. Using (30) then gives $\mu_1(x_0)_+ = 0$. Thus, $\mu_1(x)$ is continuous at $x_0$.

[45] In optimal control problems with state variable constraints, the costate variable may be discontinuous at junctions between regimes where the constraint is binding or not binding; see for instance Section 7.6 in Beavis and Dobbs (1991). Here, $\mu_1(x)$ may be discontinuous at junction points between intervals where $\widehat{I}(x) = 0$ and intervals where $\widehat{I}(x) > 0$. The proof is almost the same if the junction point is such that $\widehat{I}(x) > 0$ if $x \in (x_0 - \varepsilon, x_0]$ and $\widehat{I}(x) = 0$ if $x \in (x_0, x_0 + \varepsilon)$.

**Step 3:** $\mu_1(x) \geq 0$ *for all* $x \in [0, a]$.

Integrating $\mu_1'(x)$ given by (26) and using (28) and (29) give

$$\mu_1(0) = \int_0^a \delta(x)dx \geq 0.$$

Suppose there exist $x_0, x_1 \in [0, a]$ such that $x_0 < x_1, \mu_1(x_0) = \mu_1(x_1) = 0$ and $\mu_1(x) < 0$ if $x \in (x_0, x_1)$. Thus, from Step 2, we have $I(x) > 0$ and $\delta(x) = 0$ if $x \in (x_0, x_1)$. For $\eta_0 > 0$ small enough, we have $\mu_1'(x_0 + \eta_0) < 0$ and $\delta(x_0 + \eta_0) = 0$. Hence (26) gives

$$[\lambda - u'(R(x))]f(x) < \mu_1(x)h(x)\gamma x \frac{v'(m(x))u''(R(x))}{u'(R(x))^2}$$

for $x = x_0 + \eta_0$. The previous inequality holds when $\eta_0 \searrow 0$. Since $\mu_1(x)$ is continuous and $\mu_1(x_0) = 0$, we deduce $u'(R(x_0)) \geq \lambda$.

By a similar argument, for $\eta_1 > 0$ small enough, we have $\mu_1'(x_1 - \eta_1) > 0$ and $\delta(x_1 - \eta_1) = 0$. Thus (26) gives

$$[\lambda - u'(R(x))]f(x) > \mu_1(x)h(x)\gamma x \frac{v'(m(x))u''(R(x))}{u'(R(x))^2} > 0,$$

for $x = x_1 - \eta_1$. The previous inequality holds when $\eta_1 \searrow 0$, which implies $\lambda > u'(R(x_1))$. Thus, we have $u'(R(x_0)) \geq \lambda > u'(R(x_1))$. Since $u'' < 0$, we deduce $R(x_0) < R(x_1)$, which contradicts $R'(x) \leq 0$ and $x_0 < x_1$.

**Step 4:** $\widehat{I}'(x) \geq 0$ for all $x \in [0, a]$.

Suppose $\widehat{I}(x) > 0$ and $\widehat{I}'(x) < 0$ if $x \in [x_0, x_1] \subset (0, a]$ with $x_0 < x_1$. (6) and (8) yield $h(x) > 0$ - and thus $\varphi(x) = 0$ - and $\gamma x v'(m(x)) > u'(R(x))$ if $x \in [x_0, x_1]$. We also have $\delta(x) = 0, \mu_1(x) \geq 0$ if $x \in [x_0, x_1]$. Hence (30) gives $\varphi'(x) < 0$ if $x \in [x_0, x_1]$, which contradicts $\varphi(x) \equiv 0$ in $[x_0, x_1]$. Thus, $\widehat{I}(x)$ is non-decreasing over $[0, a]$.

**Step 5**: $\widehat{I}(x) > 0$ *for all* $x \in (0, a]$.

Step 4 implies that there exists $x_0$ in $[0, a]$ such that $\widehat{I}(x) = 0$ if $x \in [0, x_0]$ and $\widehat{I}(x) > 0$ if $x \in (x_0, a]$. Suppose $x_0 > 0$. From Step 2, we have $\mu_1(x) = 0$ for all $x \in [0, x_0]$, and

$$\mu_1(0) = \int_0^{x_0} \delta(x)dx = 0$$

implies $\delta(x) = 0$ over $[0, x_0]$.[46] (26) then gives $R'(x) = 0$ and thus $h(x) = 0$ for all $x \in [0, x_0]$. From the same argument as in Step 1, we have $m(x) = \widehat{m}(x)$, and thus $h(x) > 0$, for all $x \in [0, x_0]$, hence a contradiction.

We know from (6) and (7) that $\widehat{I}'(x) < m'(x)$ when $m'(x) > 0$, and thus Steps 1 and 5 prove Proposition 1.

Figure 8 illustrates the simulated trajectories of $\mu_1(x)$ and $\mu_2(x)$ under the calibration hypothesis introduced in Section 3.3, in the case of an exponential distribution function.

**Figure 14**

**Proof of Proposition 2**

Suppose there are $x_1, x_2, x_3$ in $[0, a]$ such that $x_1 < x_2 < x_3, h(x) = 0$ if $x \in [x_1, x_2]$ and $h(x) > 0$ if $x \in (x_2, x_3]$. Thus, $m(x)$ and $I(x)$ remain constant over $[x_1, x_2]$, and we may write $m(x) = m_0 > 0, I(x) = I_0 > 0$ and $R(x) = w - P + I_0 - m_0 = R_0$ in this interval. Let $\varphi(x)$ be defined as in the proof of Proposition 1. Using (26), (30) and $\delta(x) = h(x) = 0$ if $x \in [x_1, x_2]$ yields

$$\varphi'(x) = \lambda[1 - \frac{\gamma x v'(m_0)}{u'(R_0)}]f(x) - \gamma\mu_1(x)\frac{v'(m_0)}{u'(R_0)}, \tag{32}$$

and

$$\begin{aligned}\varphi''(x) &= \lambda[1 - \frac{\gamma x v'(m_0)}{u'(R_0)}]f'(x) - \gamma\frac{v'(m_0)}{u'(R_0)}[\lambda f(x) + \mu_1'(x)] \\ &= \lambda[1 - \frac{\gamma x v'(m_0)}{u'(R_0)}]f'(x) - \gamma\frac{v'(m_0)}{u'(R_0)}[2\lambda - u'(R_0)]f(x),\end{aligned}$$

if $x \in [x_1, x_2]$. Let

$$\Lambda(x) \equiv \frac{\varphi''(x)}{f(x)} = \lambda[1 - \frac{\gamma x v'(m_0)}{u'(R_0)}]\frac{d\ln f(x)}{dx} - \gamma\frac{v'(m_0)}{u'(R_0)}[2\lambda - u'(R_0)],$$

[46] Note that (26) and $\mu_1(x) = \mu_1'(x) = 0$ for all $x \in [0, x_0]$ imply that $\delta(x)$ is continuous in this interval.

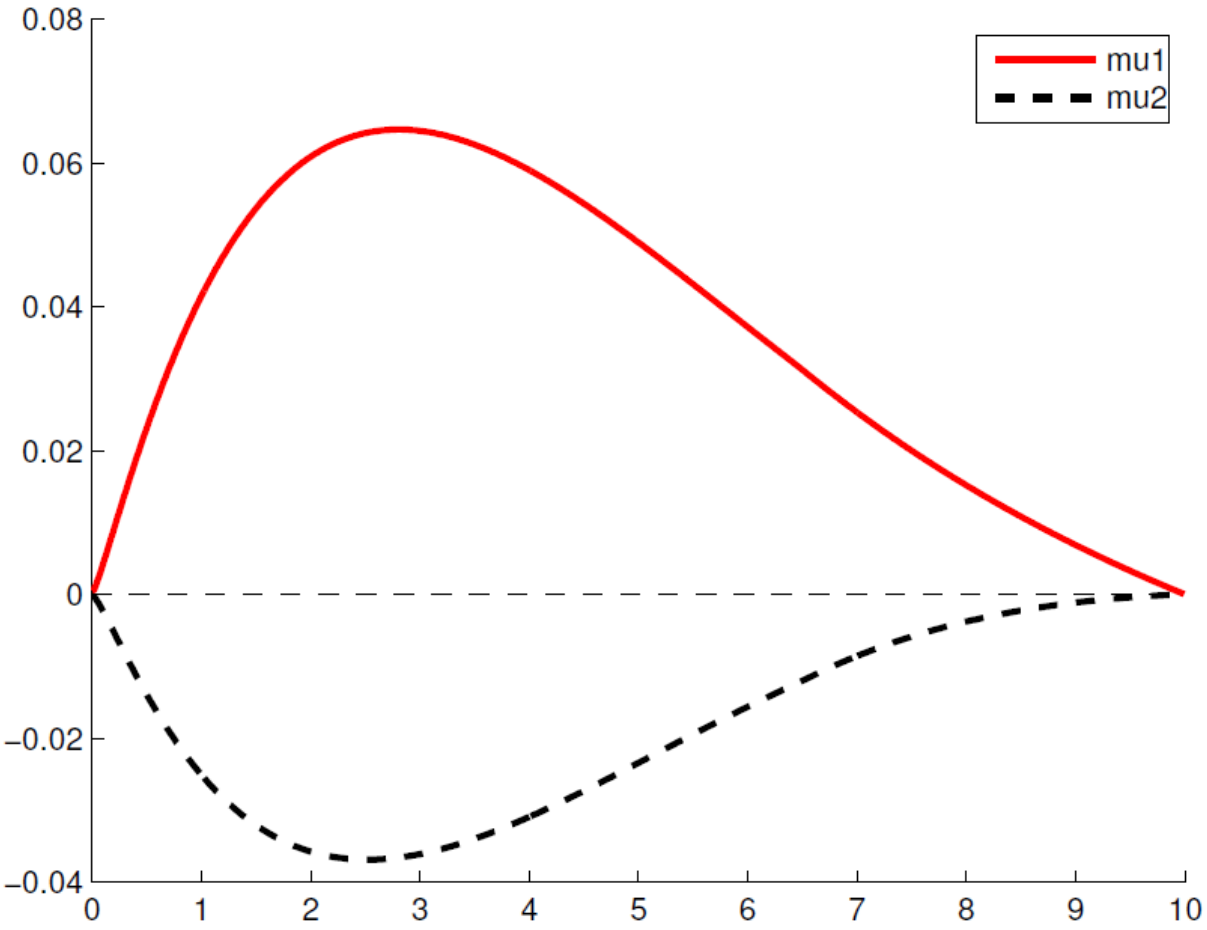



**Figure 14**

**Trajectories of co-state variables**

We have

$$\Lambda'(x) = -\lambda\gamma\frac{v'(m_0)}{u'(R_0)}\frac{d\ln f(x)}{dx} + \lambda[1 - \gamma x\frac{v'(m_0)}{u'(R_0)}]\frac{d^2\ln f(x)}{dx^2}.$$

We also have $\varphi(x) \leq 0$ if $x \in [x_1, x_2]$ and $\varphi(x_2) = 0$, which implies $\varphi'(x_2)_- \geq 0$. (30), $\delta(x_2) = 0$ and $\mu_1(x_2) > 0$[47] give $\gamma x_2 v'(m_0) \leq u'(R_0)$. If $df(x)/dx \leq 0$ and $d^2\ln f(x)/dx^2 \geq 0$, then we have $\Lambda'(x) \geq 0$ if $x \leq x_2$. Suppose there is $x_4 \in [0, x_2]$ such that $\varphi(x_4) = 0$ and $h(x) = 0$ for all $x \in [x_4, x_2]$. Since $\varphi(x) = 0$ for all $x \in [x_2, x_3]$, we have $\varphi''(x_2)_+ = 0$. Since $I_0 > 0$, $\mu_1(x)$ is differentiable at $x = x_2$. Thus, using (30) and $\delta(x) = 0$ if $x \in [x_1, x_2]$ allows us to write

$$\varphi''(x_2)_- = \varphi''(x_2)_+ + \gamma[\lambda f(x_2)x_2 + \mu_1(x_2)]\frac{d}{dx}\left(\frac{v'(m(x))}{u'(R(x))}\right)_{|x=x_2+} < 0.$$

$\Lambda(x_2)_- < 0$ and $\Lambda'(x) \geq 0$ then yield $\varphi''(x) < 0$ for all $x \in [x_4, x_2]$. Since $\varphi(x_2) = 0$ and $\varphi'(x_2)_- \geq 0$, we have $\varphi(x) < 0$ for $x < x_2$, $x$ close to $x_2$. Since $\varphi(x_2) = \varphi(x_4) = 0$, there is $x_5 \in (x_4, x_2)$ where $\varphi(x)$ has a local minimum, and thus such that $\varphi''(x_5) \geq 0$, which contradicts $\varphi''(x) < 0$ for all $x \in [x_4, x_2]$. Thus, $\varphi(x) < 0$ for all $x$ in $[0, x_2)$, which contradicts $\varphi(0) = 0$. Hence, if $h(x) > 0$ in an interval $(x_2, x_3]$, then $h(x) > 0$ in $[0, x_3]$, which shows that there exists $\overline{x} \in [0, a]$ such that $h(x) > 0$ if $x < \overline{x}$ and $h(x) = 0$ if $h(x) > \overline{x}$. We observe that $\overline{x} > 0$, for otherwise we would have $I(x) = 0$ for all $x$ in $[0, a]$.

Finally, if $x \in (0, \overline{x})$ we have $\mu_1(x) > 0, \delta(x) = 0, \varphi'(x) = 0$, and thus (30) gives $\gamma x v'(m(x)) < u'(R(x))$. Using (6) then yields $\widehat{I}'(x) > 0$.

**Proof of Corollary 1**

For notational simplicity, assume $a = 1$ and $f(x) = 1$ for all $x \in [0, 1]$. Suppose $\overline{x} < 1$. Using (30) and $h(x) = \delta(x) = 0$ if $x \in [\overline{x}, 1]$ gives

$$\varphi''(x) = -\gamma\frac{v'(\overline{m})}{u'(\overline{R})}[2\lambda - u'(\overline{R})] \equiv \overline{\varphi}''$$

[47] Step 3 in the proof of Proposition 1 shows that $\mu_1(x) > 0$ for all $x \in (0, a)$.

if $x \in (\overline{x}, 1]$. The same argument as in the proof of Proposition 2 gives $\overline{\varphi}'' = \varphi''(\overline{x})_+ < \varphi''(\overline{x})_- = 0$. Since $\varphi'(\overline{x})_+ \leq 0$, we have $\varphi'(x) < 0$ for all $x \in [\overline{x}, 1]$, which contradicts $\varphi(\overline{x}) = \varphi(1) = 0$.

**Proof of Corollary 2**

Assume $f(a) = f'(a) = 0$ and $f''(a) > 0$. Suppose $\overline{x} = a$ and thus $h(x) > 0$ for all $x \in [0, a]$.[48] We also have $\varphi'(x) = \delta(x) = 0$ for all $x$. Differentiating (30) gives

$$h(x) = -\frac{v'(m(x))J(x)}{\lambda x K(x) f(x) + v''(m(x))\mu_1(x)},$$

where

$$J(x) = -\frac{d \ln f(x)}{dx}\mu_1(x) + f(x)[2\lambda - u'(R(x))],$$
$$K(x) = v''(m(x)) + \frac{\gamma x u''(R(x))v'(m(x))^2}{u'(R(x))^2} < 0.$$

The rest of the proof is in three steps.

**Step 1**: *$J(x) > 0$ if $x \in (0, a)$ and $J(a) = J'(a) = J''(a) = h(a) = 0$.*

Using $K(x) < 0, v''(m(x)) \leq 0, \mu_1(x) > 0$ and $h(x) > 0$ gives $J(x) > 0$ if $x \in (0, a)$. Using $\mu_1(a) = f(a) = 0$ gives $J(a) = 0$. Furthermore, we have

$$J'(x) = -\frac{d \ln f(x)}{dx}\mu_1'(x) - \frac{d^2 \ln f(x)}{dx^2}\mu_1(x)$$
$$+ f'(x)[2\lambda - u'(R(x))] - f(x)u''(R(x))R'(x). \tag{33}$$

Using $\mu_1(a) = f(a) = 0, \delta(x) = 0$ for all $x$ and (26) gives $\mu_1'(a) = 0$. (33) and $d \ln f(x)/dx \nrightarrow -\infty, d^2 \ln f(x)/dx^2 \nrightarrow \pm\infty$ when $x \to a$ gives $J'(a) = 0$. Since $J(x) > 0$ if $x \in (0, a)$ and $J(a) = J'(a) = 0$, we deduce that $J(x)$ reaches a local minimum over $[0, a]$ at $x = a$, which implies $J''(a) \geq 0$.

Using L'Hôpital's rule twice yields $h(a) = -v'(m(a))J''(a)/\lambda a K(a) f''(a) = 0$. Since $h(x) \geq 0$ for all $x$, we deduce $J''(a) \leq 0$, and thus $J''(a) = h(a) = 0$.

[48] We assume w.l.o.g. that $h(x)$ is continuous at $x = a$.

**Step 2**: $u'(R(a)) = \gamma a v'(m(a)) = 2\lambda$.

Since $f(a) = f'(a) = \mu_1(a) = \mu_1'(a) = 0$, we deduce $u'(R(a)) = \gamma a v'(m(a))$ from (26) and $\varphi'(x) \equiv 0$ by using the L'Hôpital's rule twice. Furthermore, (26) gives $\mu_1''(a) = 0$ and (33) then yields $J''(a) = f''(a)[2\lambda - u'(R(a))]$, which implies $u'(R(a)) = 2\lambda$.

**Step 3**: *Let* $\xi(x) \equiv u'(R(x))\varphi'(x)$*, where* $\varphi(x)$ *is defined by (25). We have* $\xi'''(a) < 0$*, which contradicts* $\varphi(x) = 0$ *for all* $x \in [0, a]$ *when* $\overline{x} = a$.

$\overline{x} = a$ implies $\xi(x) = 0$ for all $x \in [0, a]$. We may write $\xi(x) = \lambda f(x)\Delta_1(x) - \gamma\Delta_2(x)$, with $\Delta_1(x) = u'(R(x)) - \gamma x v'(m(x)), \Delta_2(x) = \mu_1(x)v'(m(x))$. We have $\Delta_1(a) = 0, \Delta_1'(a) = -\gamma v'(m(a))$ from $h(a) = 0$ and $u'(R(a)) = \gamma a v'(m(a))$. Using (26) and Step 2 gives

$$\begin{aligned}\Delta_2'''(a) &= \mu_1'''(a)v'(m(a)) \\ &= f''(a)[\lambda - u'(R(a))]v'(m(a)) \\ &= -\lambda f''(a)v'(m(a)).\end{aligned}$$

We have

$$\begin{aligned}\xi''(x) = \lambda f''(x)\Delta_1(x) + 2\lambda f'(x)\Delta_1'(x) \\ + \lambda f(x)\Delta_1''(x) - \gamma\Delta_2''(x),\end{aligned}$$

and thus, using $\Delta_1(a) = 0$ and $f(a) = f'(a) = 0$, we may write

$$\xi'''(a) = 3\lambda f''(a)\Delta_1'(a) - \gamma\Delta_2'''(a) = -\frac{4\lambda^2 f''(a)}{a} < 0.$$

**Proof of Proposition 3**

The optimal non-linear indemnity schedule $I(m)$ is such that

$$I'(m) = \frac{\widehat{I}'(x)}{m'(x)} \text{ when } m = m(x).$$

for all $m \in (0, \overline{m})$. Thus, (6), (7), (30) and $\varphi'(x) = \delta(x) = 0$ if $x \in (0, \overline{x})$ give

$$I'(m(x)) = 1 - \frac{\gamma x v'(m(x))}{u'(R(x))} = \mu_1(x)\frac{\gamma v'(m(x))}{\lambda f(x)u'(R(x))},$$

which implies $I'(m) \in (0,1)$ for all $m \in (0, \overline{m}), I'(\overline{m}) = 0$ if $\overline{x} = a, I'(\overline{m}) > 0$ if $\overline{x} < a$, where $\overline{m} = m(\overline{x})$.

All types $x \geq \overline{x}$ choose $\overline{m} = m(\overline{x})$, and thus the optimal allocation is sustained by an indemnity schedule such that $I(m) = I(\overline{m})$ for $m \geq \overline{m}$.

Let $I'(0) = \lim_{x \to 0} I'(m) \geq 0$. The rest of the proof shows that $mv''(m)/v'(m) \to \eta \in (0,1)$ when $m \to 0$ (an assumption made in what follows) is a sufficient condition for $I'(0) > 0$. The following lemma will be an intermediary step in an a contrario reasoning.

**Lemma 5** *Suppose $I'(0) = 0$, then: (i) $h(x) \to +\infty$ when $x \to 0$. (ii) There exists a sequence $\{x_n, n \in \mathbb{N}\} \subset (0, a]$ such that $0 < x_{n+1} < x_n$ for all $n, x_n \to 0$ when $n \to \infty$ and $m(x_n)/x_n > h(x_n)$ for all $n \in \mathbb{N}$.*

**Proof of Lemma 5**

$(i)$: Note that $I'(0) = 0$ implies $C(x) \equiv xv'(m(x)) \to u'(w-P)/\gamma$ when $x \to 0$. If $(i)$ does not hold, then there exists a sequence $\{x_n, n \in \mathbb{N}\} \subset (0, a]$ such that $0 < x_{n+1} < x_n$ for all $n, x_n \to 0$ when $n \to \infty$ and $h(x_n) \to \overline{h} < +\infty$ when $n \to +\infty$. Using $v(0) = 0$ and L'Hôpital's rule yields

$$\lim_{x \to 0} C(x) = \frac{1}{\lim_{x \to 0} \left[ -\frac{v''(m(x))}{v'(m(x))^2} h(x) \right]} = \frac{1}{\eta \overline{h}} \lim_{x \to 0} \left[ m(x) v'(m(x)) \right].$$

Furthermore, $mv''(m)/v'(m) \to \eta > 0$ implies $mv'(m) \to 0$ when $m \to 0$. Hence, $C(x) \to 0$ when $x \to 0$, which contradicts $C(x) \to u'(w-P)/\gamma > 0$ when $x \to 0$.

$(ii)$: Let $x_0$ such that $h(x)$ is continuous over $(0, x_0]$ and consider the decreasing sequence $\{x_n, n \in \mathbb{N}\}$ defined by $x_n = \sup\{x \in (0, x_0] \,|\, h(x') \geq n \text{ if } x' \leq x\}$. $x_n$ is well-defined and such that $x_n \to 0$ when $n \to \infty$ from $(i)$ and, using the continuity of $h(x)$, we have $h(x_n) = n$ and $h(x) > n$ if $x < x_n$. Thus,

$$\frac{m(x_n)}{x_n} = \frac{\int_0^{x_n} h(x) dx}{x_n} > n = h(x_n),$$

which completes the proof of $(ii)$.

We are now in the position to end up the proof of the Proposition. Let us suppose $I'(0) = 0$, and let $D(x) \equiv \gamma x v'(m(x)) - u'(R(x))$ with $D(x) < 0$ if $x > 0$ from $\widehat{I}'(x) > 0$, and $D(0) = 0$ from $I'(0) = 0$. We thus have $D'(x) < 0$ for $x$ close to 0. We have

$$\begin{aligned} D'(x) &= \gamma[v'(m(x) + xv''(m(x))h(x)] - u''(R(x))R'(x) \\ &= \frac{\gamma x v'(m(x))}{m(x)} \left[ \frac{m(x)}{x} + h(x) \left( \frac{v''(m(x))m(x)}{v'(m(x))} + \frac{u''(R(x))}{u'(R(x))} m(x) \right) \right]. \end{aligned}$$

Consider the sequence $\{x_n, n \in \mathbb{N}\}$ defined in Lemma 5-$(ii)$. Using $m(x_n)/x_n > h(x_n)$ gives

$$D'(x_n) = \frac{\gamma x_n h(x_n) v'(m(x_n))}{m(x_n)} \left[ 1 + \frac{v''(m(x_n))m(x_n)}{v'(m(x_n))} + \frac{u''(R(x_n))}{u'(R(x_n))} m(x_n) \right]$$

Since $x_n \to 0$ when $n \to +\infty$, $u''(R(x))/u'(R(x)) \to u''(w-P)/u'(w-P)$ and $m(x) \to 0$ when $x \to 0$, and $-v''(m)m/v'(m) \to \eta$ when $m \to 0$, we deduce that $\eta < 1$ is a sufficient condition for $D'(x_n) > 0$ when $n$ is large enough, which is a contradiction. We deduce $I'(0) > 0$ when $\eta < 1$.

# Appendix 2

## 2-A: Computational approach

Our simulations are performed through a discretization method. Under the notations that are standard in this field, an optimal control problem is usually written as follows, by denoting $x$ the vector of state variables and $u$ the vector of controls that are function of time $t \in \mathbb{R}$:

| | |
|---|---|
| $\min \ J(x(\cdot), u(\cdot)) = g_0(t_f, x(t_f))$ | Objective (Mayer form) |
| $\dot{x}(t) = f(t, x(t), u(t)) \quad \forall t \in [0, t_f]$ | Dynamics |
| $u(t) \in U \quad$ for a.e. $t \in [0, t_f]$ | Admissible Controls |
| $g(x(t), u(t)) \leq 0$ | Path Constraints |
| $\Phi(x(0), x(t_f)) = 0$ | Boundary Conditions |

The time discretization is as follows:

$$\begin{array}{rcl} t \in [0, t_f] & \longrightarrow & t_0 = 0, \ldots, t_N = t_f \\ x(.), u(.) & \longrightarrow & X = \{x_0, \ldots, x_N, u_0, \ldots, u_N\} \\ \hline \text{Objective} & \longrightarrow & \min\ g_0(t_f, x_N) \\ \text{Dynamics} & \longrightarrow & x_{i+i} = x_i + hf(x_i, u_i) \quad i = 0, \ldots, N \\ \text{Admissible Controls} & \longrightarrow & u_i \in \mathbf{U} \quad i = 0, \ldots, N \\ \text{Path Constraints} & \longrightarrow & g(x_i, u_i) \leq 0 \quad i = 0, \ldots, N \\ \text{Boundary Conditions} & \longrightarrow & \Phi(x_0, x_N) = 0 \end{array}$$

We therefore obtain a nonlinear programming problem on the discretized state and control variables. In BOCOP, the discretized nonlinear optimization problem is solved by the Ipopt solver that implements a primal-dual interior point algorithm; see Wachter and Biegler (2006). The derivatives required for the optimization are computed by the automatic differentiation tool Adol-C; see Walther and Griewank (2012).

## 2-B: Complementary proofs

**Proof of Lemma 2**

Let $\widehat{I}(x)$, $x \in [0, x^*]$, $P$ and $c^*$ be given, with $I^* = \widehat{I}(x^*), m^* = m(x^*)$ and $I^* \leq m^*$. Consider the sub-problem in which $\{\widehat{I}(x), m(x), g(x), h(x),\ x \in [x^*, a]\}$ maximizes

$$\int_{x^*}^{a} \Big\{ u(w - P + \widehat{I}(x) - m(x)) + h_0 - \gamma x[1 - v(m(x))] \Big\} f(x) dx, \tag{34}$$

subject to (7) and (10)-(12).

Let $\mu_1(x)$ and $\mu_2(x)$ be co-state variables respectively for $\widehat{I}(x)$ and $m(x)$ and let $\eta(x)$, and $\lambda$ be Lagrange multipliers respectively for (11) and (12) in this sub-problem.[49] The Hamiltonian is written as

$$\begin{aligned} \mathcal{H} = {} & [u(R(x)) + \gamma x v(m(x))] f(x) + [\mu_1(x) - \eta(x)] g(x) \\ & + [\mu_2(x) + \eta(x)] h(x) - \lambda [\widehat{I}(x) + c] f(x), \end{aligned}$$

[49] We can straightforwardly check that (8) is not binding in this sub-problem.

and the optimality conditions are

$$\mu_1(x) - \eta(x) \leq 0, = 0 \text{ if } g(x) > 0, \tag{35}$$

$$\mu_2(x) + \eta(x) = 0, \tag{36}$$

$$\mu_1'(x) = [\lambda - u'(R(x))]f(x), \tag{37}$$

$$\mu_2'(x) = [u'(R(x)) - \gamma x v'(m(x))]f(x)\ , \tag{38}$$

for all $x$, with the transversality conditions $\mu_1(a) = \mu_2(a) = 0$, and $\eta(x) \geq 0$ for all $x$ and $\eta(x) = 0$ if $h(x) > g(x)$.

Let us consider $x_0 \in [x^*, a]$ such that $g(x) > 0$ if $x$ is in a neighbourhood $\mathcal{V}$ of $x_0$. Suppose $h(x) > g(x)$, and thus $\eta(x) = 0$ if $x \in \mathcal{V}$. (35) gives $\mu_1(x) = 0$, and thus $\mu_1'(x) = 0$ for all $x \in \mathcal{V}$. Then (37) gives $u'(R(x)) = \lambda$, and thus $R(x) = w - P - m(x) + \widehat{I}(x)$ is constant in $\mathcal{V}$. This implies $m'(x) - \widehat{I}'(x) = h(x) - g(x) = 0$, which contradicts $h(x) > g(x)$. We deduce that $h(x) = g(x)$ if $x \in \mathcal{V}$. (35) and (36) yield $\mu_1(x) = -\mu_2(x) = \eta(x)$, and thus $\mu_1'(x) = -\mu_2'(x)$, for all $x \in \mathcal{V}$. (37) and (38) then imply $\gamma x v'(m(x)) = \lambda$ for all $x \in \mathcal{V}$, which gives $m'(x) = -v'(m(x))/xv''(m(x))$.

Let $x_0, x_1, x_2 \in [x^*, a]$ such that $x_0 < x_1 < x_2$ with $g(x) = 0$ if $x \in [x_0, x_1]$ and $g(x) > 0$ if $x \in (x_1, x_2]$. Let us show that we cannot have $g(x) > 0$ if $x \in [x_3, x_0]$ with $x_3 < x_0$. We have $\mu_1(x) + \mu_2(x) \leq 0$ if $x \in [x_0, x_1)$ and $\mu_1(x) + \mu_2(x) = 0$ if $x \in [x_1, x_2]$. Let $\Psi(x) \equiv [\mu_1'(x) + \mu_2'(x)]/f(x)$, with $\Psi(x_1) = 0$ because $\mu_1(x) + \mu_2(x)$ reaches a local maximum at $x = x_1$. Note that $\Psi(x)$ is differentiable. Let $x \in [x_0, x_1)$. If $m'(x) = 0$ (and thus $R'(x) = 0$), we have $d[\mu_1'(x)/f(x)]/dx = 0$ and $d[\mu_2'(x)/f(x)]/dx = -\gamma v'(m(x_1)) < 0$, and thus $\Psi'(x) < 0$. If $m'(x) > 0$ (and thus $R'(x) < 0$), we have $\eta(x) = \mu_2(x) = \mu_2'(x) = 0$ and $d[\mu_1'(x)/f(x)] = -u''(R(x))R'(x) < 0$, and thus we still have $\Psi'(x) < 0$. Suppose $g(x) > 0$ if $x \in [x_3, x_0]$ with $x_3 < x_0$. In that case we would have $\mu_1(x) + \mu_2(x) = 0$ if $x \in [x_3, x_0]$, and since $\mu_1(x) + \mu_2(x) \leq 0$ if $x \in [x_0, x_1)$, we would have $\Psi(x_0) = 0$. This contradicts $\Psi(x_1) = 0, \Psi'(x) < 0$ if $x \in [x_0, x_1)$.

Suppose there are $x_0, x_1, x_2 \in [x^*, a]$ such that $x_0 < x_1 < x_2$ with $g(x) > 0$ if

$x \in [x_0, x_1]$ and $g(x) = 0$ if $x \in (x_1, x_2]$. In that case $\mu_1(x) + \mu_2(x) = 0$ if $x \in [x_0, x_1]$ and $\mu_1(x) + \mu_2(x) \leq 0$ if $x \in [x_1, x_2]$. Since $\mu_1(a) + \mu_2(a) = 0$ and $\mu_1(x)$ and $\mu_2(x)$ are continuous, we may choose $x_2$ such that $\mu_1(x_2) + \mu_2(x_2) = 0$. The same calculation as above implies $\Psi(x_1) = 0$, $\Psi'(x) < 0$ if $x \in [x_1, x_2]$ and thus $\Psi(x) < 0$ if $x \in [x_1, x_2]$, which contradicts $\mu_1(x_2) + \mu_2(x_2) = 0$.

Overall, we deduce that there exists $\widehat{x} \in [x^*, a]$ such that $\widehat{I}'(x) = 0$ if $x \in [x^*, \widehat{x}]$ and $\widehat{I}'(x) = m'(x) > 0$ if $x \in [\widehat{x}, a]$. The same reasoning - replacing $\Psi(x)$ by $\Phi(x) \equiv \mu_2'(x)/f(x)$ - shows that there exists $\widetilde{x} \in [x^*, \widehat{x}]$ such that $m'(x) = 0$, and thus $m(x) = m^*$, if $x \in [x^*, \widetilde{x}]$ and $m'(x) > 0$ if $x \in [\widetilde{x}, \widehat{x}]$. When $m'(x) > 0$, we have $\eta(x) = \mu_2(x) = 0$, and thus $\mu_2'(x) \equiv 0$ if $x \in [\widetilde{x}, \widehat{x}]$, which gives $u'(w - P - m(x) + I^*) = \gamma x v'(m(x))$, and thus $m'(x) \equiv -\gamma v'(m(x))/[\gamma x v''(m(x)) + u''(w - P - m(x) + I^*)]$. When $m'(x) = 0$, we have $\Phi'(x) < 0$ if $[x^*, \widetilde{x})$ and $\Phi'(\widetilde{x}) = 0$, and thus $\widetilde{x}$ is given by $u'(w - P - m^* + I^*) = \gamma \widetilde{x} v'(m^*)$ if $u'(w - P - m^* + I^*) > \gamma x^* v'(m^*)$, and $\widetilde{x} = x^*$ if $u'(w - P - m^* + I^*) = \gamma x^* v'(m^*)$.

If $x^* < \widehat{x}$, then replacing $m^*$ by $\widehat{m} \equiv m(\widehat{x}) > m^*$ implements the same allocation with lower audit costs. Indeed, $m(x)$ is an optimal choice of type $x$ individuals if $x > \widehat{x}$, because such individuals would prefer choosing $\widehat{m}$ rather than any $m \in [0, \widehat{m})$, and furthermore, for such individuals, there is full coverage at the margin in $(\widehat{m}, m(x)]$ and they cannot choose expenses larger than $m(x)$. In addition, the expected audit cost decreases from $c[1 - F(x^*)]$ to $c[1 - F(\widehat{x})]$ when $\widehat{m}$ is substituted for $m^*$. Thus, an optimal allocation is necessarily such that $x^* = \widehat{x}$.

**Proof of Proposition 4**

Let $\mu_1(x)$ and $\mu_2(x)$ be costate variables respectively for $\widehat{I}(x)$ and $m(x)$ and let $\delta(x)$ and $\lambda$ be Lagrange multipliers respectively for (9) and (20). The Hamiltonian is written as in the proof of Proposition 1, and the optimality conditions (25), (26) and (27) still hold. We also have $\delta(x) \geq 0$ and $\delta(x) = 0$ if $\widehat{I}(x) > 0$, and $\mu_1(x^*) + \mu_2(x^*) = 0$ from the characterization of the optimal continuation allocation. The optimality conditions

on $m^*, I^*, x^*, P$ and $A$ are written as

$$V_1' - \mu_2(x^*) = 0, \tag{39}$$

$$V_2' - \mu_1(x^*) = 0, \tag{40}$$

$$V_3' + \{u(R^*) + h_0 - \gamma x^*[1 - v(m^*)]\}f(x^*)$$
$$-\mu_1(x^*)\frac{\gamma x^* v'(m^*)}{u'(R^*)} - [\lambda - \delta(x^*)]I^* \leq 0, = 0 \text{ if } x^* > 0, \tag{41}$$

$$V_4' - \int_0^{x^*} \left[u'(R(x))f(x) + \mu_1(x)h(x)\gamma x\frac{v'(m(x))u''(R(x))}{u'(R(x))^2}\right] dx = 0, \tag{42}$$

$$V_5' + \lambda = 0, \tag{43}$$

respectively, where $V_1', V_2', ...$ denote the partial derivatives of $V(m^*, I^*, x^*, P, A)$ and $R^* \equiv R(x^*) = w - P - m^* + I^*$. Define $\varphi(x)$ for all $x \in [0, x^*]$ by (25) as in the proof of Proposition 1.

**Step 1**: $m(x) > 0$ for all $x > 0$.

Identical to Step 1 in the proof of Proposition 1.

**Step 2:** *$\mu_1(x)$ is continuous in $[0, x^*]$ with $\mu_1(x) = 0$ for all $x \in [0, x^*]$* such that $\widehat{I}(x) = 0$.

Identical to Step 2 in the proof of Proposition 1.

**Step 3**: *$\mu_1(x) \geq 0$ for all $x \in [0, x^*]$* with $\mu_1(x^*) > 0$.

We know from Lemma 4 that $R(x) = w - P - m^* + I^*$ and

$$m(x) = m^* + \int_{x^*}^{x} \frac{v'(m(t))}{tv''(m(t))}dt,$$

for all $x \in [x^*, a]$. Thus,

$$V_2' = u'(w - P - m^* + I^*)[1 - F(x^*)],$$

and (40) gives $\mu_1(x^*) > 0$. The remaining part of Step 3 is the same as in the proof of Proposition 1.

**Step 4:** *$\widehat{I}(x) > 0$ for all $x \in (0, x^*]$.*

Identical to Steps 4 and 5 in the proof of Proposition 1.

**Step 5**: $x^* > 0$.

We have

$$V_3' = -\{u(R^*) + h_0 - \gamma x^*[1 - v(m^*)] + \lambda(I^* + c)\}f(x^*),$$

from the definition of $V(.)$. Thus (41) and $\delta(x^*) = 0$ give

$$\lambda c f(x^*) - \mu_1(x^*)\frac{\gamma x^* v'(m^*)}{u'(R^*)} \leq 0, = 0 \text{ if } x^* > 0,$$

which implies $x^* > 0$.

**Step 6**: *There is* $\overline{x} \in (0, x^*]$ *such that*

$$\widehat{I}'(x) > 0, h(x) = m'(x) > 0 \quad \textit{if} \quad 0 < x < \overline{x},$$
$$\widehat{I}(x) = \widehat{I}(\overline{x}), m(x) = m(\overline{x}), h(x) = 0 \quad \textit{if} \quad \overline{x} < x \leq x^*,$$
$$\widehat{I}'(0) = 0, \widehat{I}'(\overline{x}) = 0 \textit{ if } \overline{x} = a \textit{ and } \widehat{I}'(\overline{x}) > 0 \text{ if } \overline{x} < x^*.$$

Identical to the proof of Proposition 2.

Finally, $\mu_1(x^*) > 0$ shows that there is an upward discontinuity in $m(x)$ and $\widehat{I}(x)$ at $x = x^*$.

**Proof of Proposition 5**

Using $x^* > 0$ and $m'(x) > 0$ if $x \in (0, \overline{x})$ gives $m^* > 0$. The remaining part of the Proposition is a straightforward adaptation of Proposition 3.

**Proof of Lemma 3**

Similar to Lemma 1, with straightforward adaptation.

**Proof of Lemma 4**

We now have

$$V(x, \widetilde{x}) = U\left(w - P + \widehat{I}(\widetilde{x}) - m(\widetilde{x})), h_0 - \gamma x(1 - v(m(\widetilde{x}))\right).$$

A straightforward adaptation of the proof of Lemma 1 shows that (17) is a necessary condition for incentive compatibility. (17) gives

$$\frac{\partial V(x,\widetilde{x})}{\partial \widetilde{x}} = \gamma v'(m(\widetilde{x}))m'(\widetilde{x})U'_H(R(\widetilde{x}),H(x,\widetilde{x}))\left[x - \widetilde{x}A(x,\widetilde{x})\right],$$

where

$$H(x,\widetilde{x}) \equiv h_0 - \gamma x(1 - v(m(\widetilde{x})), H(\widetilde{x},\widetilde{x}) \equiv H(\widetilde{x}),$$
$$A(x,\widetilde{x}) \equiv \frac{U'_R(R(\widetilde{x}),H(x,\widetilde{x}))U'_H(R(\widetilde{x}),H(\widetilde{x}))}{U'_R(R(\widetilde{x}),H(\widetilde{x}))U'_H(R(\widetilde{x}),H(x,\widetilde{x}))}.$$

Using $U''_{H^2} < 0$ and $U''_{RH} > 0$ gives $A(x,\widetilde{x}) > 1$ if $\widetilde{x} > x$ and $A(x,\widetilde{x}) < 1$ if $\widetilde{x} < x$, with $A'_{\widetilde{x}}(x,\widetilde{x})_{|\widetilde{x}=x} > 0$, and thus[50]

$$\frac{\partial^2 V(x,\widetilde{x})}{\partial \widetilde{x}^2}\,|_{\widetilde{x}=x} = -\gamma v'(m(x))m'(x)U'_H(R(x),H(x))[1 + A'_{\widetilde{x}}(x,\widetilde{x})_{|\widetilde{x}=x}].$$

Thus incentive compatibility gives (18). Conversely, assume that (17) and (18) hold. We have

$$\frac{\partial V(x,\widetilde{x})}{\partial \widetilde{x}} \leq \gamma v'(m(\widetilde{x}))m'(\widetilde{x})U'_H(R(\widetilde{x}),H(x,\widetilde{x}))(x - \widetilde{x}) < 0 \text{ if } \widetilde{x} > x,$$
$$\frac{\partial V(x,\widetilde{x})}{\partial \widetilde{x}} \geq \gamma v'(m(\widetilde{x}))m'(\widetilde{x})U'_H(R(\widetilde{x}),H(x,\widetilde{x}))(x - \widetilde{x}) > 0 \text{ if } \widetilde{x} < x,$$

which implies incentive compatibility.

**Proof of Proposition 6**

The notations of costate variables and Lagrange multipliers are the same as in the proof of Proposition 1. Observe first that Steps 1-4 of this proof remain valid, with an unchanged definition of $\varphi(x)$, just replacing (30) by

$$\varphi'(x) = [\lambda(1+\sigma)f(x) - \delta(x)]\left[1 - \frac{\gamma x v'(m(x))}{u'(R(x))}\right] - \gamma\mu_1(x)\frac{v'(m(x))}{u'(R(x))}. \qquad (44)$$

and $\lambda$ by $\lambda(1+\sigma)$ in (26).

[50] On can check that $A'_{\widetilde{x}|\widetilde{x}=x} > 0$ if $U'_H U''_{RH} - U'_R U''_{H^2} > 0$, which holds when $U''_{RH} > 0, U''_{H^2} < 0$ as postulated, but which is also compatible with $U''_{RH} < 0$.

Suppose that $\widehat{I}'(x) > 0$ if $x < \varepsilon$, with $\varepsilon > 0$. Hence $\widehat{I}(x) > 0$ (and thus $\delta(x) = 0$) for all $x > 0$. Using (6) gives

$$h(x) > 0, \tag{45}$$

$$1 - \frac{\gamma x v'(m(x))}{u'(R(x))} > 0, \tag{46}$$

if $x < \varepsilon$. (45) implies $\varphi(x) = \varphi'(x) = 0$ if $x < \varepsilon$. Furthermore, using (26) (in which $\lambda$ is replaced by $\lambda(1+\sigma)$), (29) and $\mu_1(a) = 0$ yields

$$\mu_1(0) = -\int_{\underline{x}}^{a} \mu_1'(x)dx = \int_0^a \delta(x)dx - \lambda\sigma = -\lambda\sigma < 0,$$

and thus $\mu_1(x) < 0$ for $x$ small enough. (44) and (46) then yield $\varphi'(x) > 0$, hence a contradiction. Since we know from Step 4 that $\widehat{I}(x)$ is non-decreasing, we deduce that there exists $d > 0$ such that $\widehat{I}(x) = 0$ if $x \leq d$ and $\widehat{I}(x) > 0$ if $x > d$.

The simulated trajectories of $\mu_1(x)$ and $\mu_2(x)$ are illustrated in Figure 9 in the case of an exponential distribution function, with $\sigma = 0.1$ and with the same calibration as in Section 3.4. We have $\mu_1(x) = \mu_2(x) = 0$ when $x \leq d$ and $\mu_1(x) > 0, \mu_2(x) < 0$ when $x > d$, with $d \simeq 0.41$.

The characterization of the indemnity schedule $I(m)$ is derived in the same way as in Proposition 3, with $D = m(d)$.[51]

**Figure 15**

**Proof of Corollary 3**

Similar to Corollary 1.

**Proof of Corollary 4**

Similar to Corollary 2.

[51] Note however, that we may have $I'(D_+) = 0$ as illustrated in Figure 6 (bottom) and 7.

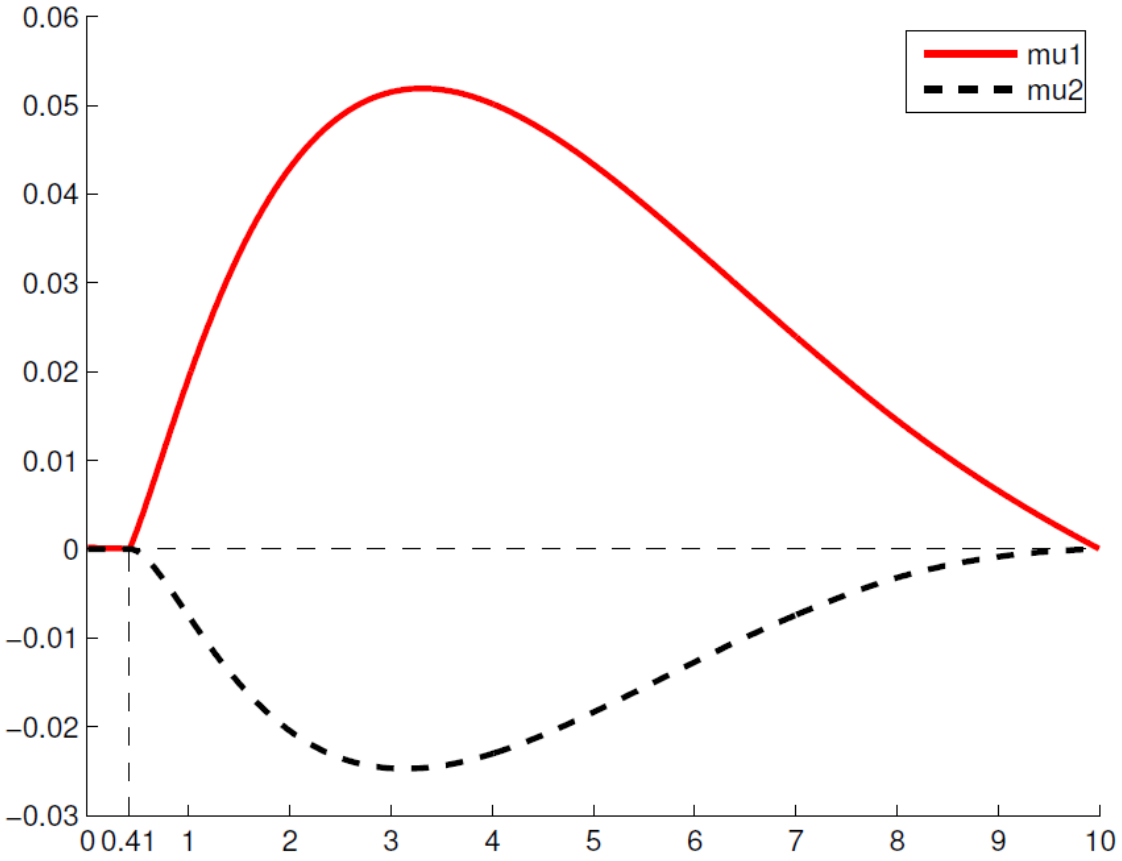



**Figure 15**

**Co-state variables under loading**

# References


Arrow, K.J., 1963, "Uncertainty and the welfare economics of medical care", *American Economic Review*, 53, 941-973.

Arrow, K.J., 1968, "The economics of moral hazard: further comment", *American Economic Review*, 58, 537-539.

Arrow, K.J., 1971, *Essays in the Theory of Risk Bearing*, Markham Publishing, Chicago.

Arrow, K.J., 1976, "Welfare analysis of changes in health co-insurance rates", in *The Role of Health Insurance in the Health Services Sector*, R. Rosett (ed.), NBER, New York, 3-23.

Beavis, B., and I. Dobbs, 1991, *Optimization and Stability Theory for Economic Analysis*, Cambridge University Press, Cambridge.

Betts, J.T., 2001, *Practical Methods for Optimal Control Using Nonlinear Programming*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Blomqvist, A., 1997, "Optimal non-linear health insurance", *Journal of Health Economics*, 16, 303-321.

Bond, E, and K.J. Crocker, "Hardball and the soft touch: the economics of optimal insurance contracts with costly state verification and endogenous monitoring costs", *Journal of Public Economics*, 63, 239-264.

Bonnans, F., Giorgi, D., Grelard, V., Heymann, B., Maindrault, S., Martinon, P., and O. Tissot, 2016, *Bocop - A Collection of Examples*, Technical Report, INRIA.

Cutler, D.M., and R.J. Zeckhauser, 2000, "The anatomy of health insurance", in *Handbook of Health Economics, vol.1*, A. Culyer, and J.P. Newhouse (Eds), North-Holland: Amsterdam, 563-643.

Drèze, J.H. and E. Schokkaert, 2013, "Arrow's theorem of the deductible: moral hazard and stop-loss in health insurance", *Journal of Risk and Uncertainty*, 47(2), 147-163.

Ebert, U., 1992, "A reexamination of the optimal nonlinear income tax", *Journal*

*of Public Economics*, 49, 47-73.

Ellis, R.P., S. Jiang, and W.G. Manning, "Optimal health insurance for multiple goods and time periods", *Journal of Health Economics*, 41, 89-106.

Evans, W.N. and W.K. Viscusi, 1991, "Estimation of state dependent utility functions using survey data", *Review of Economics and Statistics*, 73, 94-104.

Feldman, R. and B. Dowd, 1991, "A new estimate of the welfare loss of excess health insurance", *American Economic Review*, 81, 297-301.

Feldstein, M., 1973, "The welfare loss of excess health insurance", *Journal of Political Economy*, 81, 251-280.

Feldstein, M. and B. Friedman, 1977, "Tax subsidies, the rational demand for insurance and the health care crisis", *Journal of Public Economics*, 7, 155-178.

Finkelstein, A., Luttmer, E.F.P., and M.J. 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, 221-258.

Gollier, C., 1987, "Pareto-optimal risk sharing with fixed cost per claim", *Scandinavian Actuarial Journal*, 13, 62-73.

Holmström, B., 1979, "Moral hazard and observability", *Bell Journal of Economics*, 10, 74-91.

Huberman, G., D. Mayers and C.W. Smith, Jr., 1983, "Optimum insurance policy indemnity schedules", *Bell Journal of Economics*, 14, 415-426.

Kaiser Family Foundation (2009), *Cost Sharing for Health Care: France, Germany, and Switzerland*, The Henry J. Kaiser Family Foundation, January 2009.

Laffont, J.-J. and J.-C. Rochet, 1998, "Regulation of a risk-averse firm", *Games and Economic Behavior*, 25, 149-173.

Lollivier, S. and J.-C. Rochet, 1983, "Bunching and second-order conditions: A note on optimal tax theory", *Journal of Economic Theory*, 31, 2, 392-400.

Nocedal, J. and S.J. Wright, 1999, *Numerical Optimization*, Springer-Verlag, New-York.

Pauly, M., 1968, "The economics of moral hazard: comment", *American Economic Review*, 58, 531-537.

Pflum, K.E., 2015, "Physician incentives and treatment choices", *Journal of Economics and Management Strategy*, 24, 712-751.

Picard, P., 2000, "On the design of optimal insurance policies under manipulation of audit cost", *International Economic Review*, 41, 4, 1049-1071.

Picard, P., 2013, "Economic analysis of insurance fraud", in *Handbook of Insurance*, G. Dionne (Ed), Second Edition, Springer, 349, 395.

Picard, P., 2016, "A note on health insurance under ex post moral hazard", *Risks*, 4, 38, 1-9.

Raviv, A., 1979, "The design of an optimal insurance policy", *American Economic Review*, 69, 854-896.

Rees, R. and A. Wambach, 2008, *The Microeconomics of Insurance*, in *Foundations and Trends in Microeconomics*, vol.4:1-2, 1-163.

Salanié B., 1990, "Sélection adverse et aversion pour le risque", *Annales d'Economie et de Statistiques*, 18, 131-150.

Townsend, R., 1979, "Optimal contracts and competitive markets with costly state verification", *Journal of Economic Theory*, 21, 265-293

Viscusi, W.K. and W.N. Evans, 1990, "Utility functions that depend on health status: estimates and economic implications", *American Economic Review*, 80, 353-374.

Wächter, A. and L.T. Biegler, 2006, "On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming", *Mathematical Programming*, 106, 1, 25-57.

Walther, A. and A. Griewank, 2012, "Getting started with adol-c", in *Combinatorial Scientific Computing,* U. Naumann and O. Schenk, (Eds), Chapman-Hall CRC Computational Science.

Weymark, J.A., 1986, "A reduced-form optimal nonlinear income tax problem",

*Journal of Public Economics*, 30, 2, 199-217.

Winter, R.A., 2013, "Optimal insurance contracts under moral hazard", in *Handbook of Insurance*, G. Dionne (Ed), Second Edition, Springer, 205-230.

Zeckhauser, R., 1970, "Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives", *Journal of Economic Theory*, 2, 10-26.