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

Pierre Martinon, Pierre Picard, Anasuya Raj

ECOLE POLYTECHNIQUE

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

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

Pierre MARTINON
Pierre PICARD
Anasuya RAJ

July 1st, 2016

Cahier n° 2016-09

DEPARTEMENT D'ECONOMIE
Route de Saclay
91128 PALAISEAU CEDEX
(33) 1 69333033
http://www.portail.polytechnique.edu/economie/fr
mariame.seydi@polytechnique.edu

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

Pierre Martinon*, Pierre Picard† and Anasuya Raj‡

July 1st, 2016

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 on their health expenditures. We characterize the trade-off between ex ante risk sharing and ex post incentive compatibility, in an optimal revelation mechanism under hidden information and risk aversion. We establish that the optimal contract provides partial insurance at the margin, with a deductible when insurers' rates are affected by a positive loading, and that it may also include an upper limit on coverage. We show that 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.
†Ecole Polytechnique, Department of Economics, France.
‡Ecole Polytechnique, Department of Economics, France.

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 have before us.

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.[1] We will analyze this problem with the double concerns of robustness

[1] Winter (2013) surveys the literature on insurance under ex ante and ex post moral hazard. The ex post moral hazard information problem was identified by Zeckhauser (1970) and addressed firstly by Blomqvist (1997). The latter argues that the indemnity schedule is S-shaped, with marginal coverage increasing for small expenses and decreasing for large expenses. Unfortunately, he overlooks

of theoretical conclusions and, as far as possible, a conformity with economic reality.

Not surprisingly, under ex post moral hazard, the trade-off between incentives and risk sharing leads to a partial coverage at the margin. However, Salanié (1990) and Laffont and Rochet (1998) have shown that bunching (i.e., the incomplete separation of agents according to their type) frequently arises in adverse selection principal-agent models with risk averse agents. This may also be the case in the Mirlees' optimal income tax model, as shown by Lollivier and Rochet (1983) and Weymark (1986). In our insurance setting, the optimal contract actually does not always fully separate individuals according to their health state. Under specific assumptions about the probability distribution of health states, that will correspond to a cap on health expenses and insurance indemnities that is reached by a non-negligible fraction of policyholders.[2] Thus, the optimal contract specifies a partial reimbursement at the margin, with bunching "at the top". Furthermore, a deductible may be optimal 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:

important technical aspects (including bunching and limit conditions), which considerably reduces the relevance of his conclusions. All in all, as shown in this paper, in most cases, the optimal indemnity schedule is in fact not S-shaped. Drèze and Schokkaert (2013) show that the Arrow's theorem of the deductible extends to the case of ex post moral hazard. However, they directly postulate that the insurance premium is computed with a positive loading factor, presumably because of transaction costs. They do not address the question of whether ex post and ex ante moral hazard differ in this respect, independently of the existence of transaction costs.

[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.

[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). 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).

in the absence of transaction costs, a deductible may be optimal under ex ante moral hazard (Holmström, 1979), but not under ex post moral hazard. We will show that this characterization is robust to changes in the modelling, including the case where income is affected by a background risk and the one where preferences are not separable between wealth and health. Finally, we will show how the "umbrella policy" approach - in which insurance indemnity depends on total medical expenses, both of them being measured on a per period basis - can be connected to the "fee-for-service" approach, in which medical services are unbundled, and separately paid for by patients and insurers.

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.[4] 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 practioners or their dentists to cure benign pathologies, while insurers have control over more serious expenses, in particular surgeries or other types of hospital care.

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 a limit on out-of-pocket expenditures in the indemnity schedule.

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

In brief, the objective and results of this paper are twofold. Firstly, we characterize the optimal health insurance indemnity schedule under ex post moral hazard in a way which is as robust as possible - a task that does not seem to have been performed in a satisfactory way sofar -, and secondly, we extend this analysis to a costly state verification setting. To do so, we proceed as follows. 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 technics, and they are also solved through a computational approach. Section 4 appraises the robustness of our results by considering alternative models, with correlated background risk, with non-separable utility, and with insurance loading, respectively. Section 5 immerses the ex post moral hazard problem in a costly state verification setting, where health expenses may be audited. Section 6 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 section, as in Blomqvist (1997), we will start by restricting 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.[5] The non-separability case will be considered in Section 4. 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

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

care, including insurance transfers.

Health may be negatively affected by illness, but it increases with the 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) and $m \geq 0$ denotes medical 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 $x \in [0, a)$.[6]

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.[7] Function $I(.) : \mathbb{R}_+ \to \mathbb{R}_+$ is supposed to be continuous, nondecreasing, piecewise continuously differentiable and such that $I(0) = 0$.[8] We have $T = m + P - I(m)$ and $R = w - T =$

[6] 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.

[7] This is an umbrella policy in which expenses and indemnity are evaluated on a per period basis. The link with a fee-for-service approach (with unbundled compensation rules) will be considered in subsection 3.4.

[8] In addition to being realistic, assuming that $I(m)$ is nondecreasing is not a loss of generality if policyholders can claim insurance payment for only a part of their medical expenses: in that case,

$w - P - m + I(m)$.[9] A type x individual chooses her health care expenses $m(x)$ in 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

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

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 $I(0) = 0$.

[9] 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 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.

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

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

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.

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 continous 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 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.[10] 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*

$$\frac{d \ln f(x)}{dx} < x \frac{d^2 \ln f(x)}{dx^2} \quad \textit{for all } x \in [0, a]. \tag{10}$$

Then there is $\overline{x} \in (0, a]$ *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* $\overline{x} < a$.

[10] Note that $\widehat{I}(x)$ and $m(x)$ are piecewise differentiable because $I(m)$ is piecewise differentiable. This allows us to use the 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.

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

The optimal contract trades off risk-sharing and incentives to not overspend for medical services. Under condition (10), this trade-off may tip in favor of the incentive effect when x is large enough.[11] 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 capping expenditures and indemnities to be suboptimal, 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.

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

[11] When $f(x)$ is non-increasing, a sufficient condition for (10) to hold is written as $d^2 \ln f(x)/dx^2 \geq 0$, i.e. $\ln f(x)$ is non-increasing and weakly convex. This is the case, for instance, if the distribution of x is uniform or exponential.

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, which corresponds to the case $u'(\overline{R}) - \gamma \overline{x} v'(\overline{m}) = 0$, 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 $I(\overline{m})$.[12] 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.

[12] 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 it may be low.

3.3 Simulation

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 dynamics equations.[13] We assume that x is distributed over $[0, 10]$, i.e., $a = 10$, either exponentially, i.e., $f(x) = \lambda e^{-\lambda x} + e^{-\lambda a}/a$, with $\lambda = 0.25$,[14] 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 the (m, I) space for $x \in \{0.3, 7, 9\}$ when x is uniformly distributed. Parameters σ 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 $m(10) \simeq 0.772$ and $\widehat{I}(x) = I(m(x))$ goes from $I(0) = 0$ to $I(0.772) \simeq 0.457$, 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 $I'(0.772) = 0$.

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 $\overline{m} \simeq 0.872$. $I(m)$ has an angular point at $m = \overline{m}$. Figure 2 illustrates the case of types $x = 7$ and $x = 9$: in both cases, the

[13] Such direct approach methods are usually less precise than indirect methods based on Pontryagin's Maximum Principle, but they are more robust with respect to the initialization. We here use the Bocop software (see http://itn-sadco.inria.fr/software/bocop-software). 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.

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

optimal expenses are equal to $\overline{m}$. As in Figure 1, we have $I'(0) > 0$: the optimal indemnity schedule does not include a deductible.

Figures 1 and 2

3.4 Fee-for-service payment

In most health insurance contracts, medical services (visits to doctors, drug purchases, hospital stays...) are unbundled, and separately paid for by patients and insurers. So far, we have defined the indemnity schedule as an "umbrella policy" that gives the total insurance idemnity I as a function of the total medical expenses m, irrespective of the selection of medical services that are at the origin of these expenses. Let us sketch a more general model in order to connect the two approaches.

Assume that the health state is denoted by $\omega = (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 = \Omega$, where ω_1 is publicly observable and ω_2 is private information of the policyholder. Ω_1 and Ω_2 are multi-dimensional sets that include all possible diseases, characterized by the type and severity of pathologies, with one of these states corresponding to good health. Assume also that there exists a function $g(.) : \Omega \longrightarrow [0, a]$ such that $x = g(\omega)$ is the severity of illness in state ω. As before, severity refers to the cost of medical services that are required to reach a given utility gain through health improvement. The probability distribution of x is induced by a probability measure $\mathbb{P}_\Omega(.)$ over Ω.

Let us index medical services by $s = 1, ..., S$, and let p_s denote the market price for a unit quantity of service s. The total medical expenses of a patient who benefits n_s times from y_s units of type s service is $m = \sum_{s=1}^{S} p_s n_s y_s$.[15] Let $n = (n_1, ..., n_S)$ and $y = (y_1, ..., y_S)$. Type ω individuals choose (n, y) in a type-dependent set $\mathcal{K}(\omega_1) \subset \mathbb{N}^S \times \mathbb{R}_+^S$

[15] For instance, if s corresponds to primary health care provided by general practionners, then n_s is the number of visits to such doctors during the policy year. The policyholder purchases a more or less important quantity of each type of medical services. For instance, he may visit a doctor who spends more time with his patients or who has better skills or more experience, and this doctor would charge a higher price per visit.

that includes the medical services that can be helpful in state ω. We thus assume that the set of useful medical services depends on the observable component ω_1, while the severity of the desease depends on ω_1 and ω_2 and is thus private information to the policyholder. The policyholder's utility is still written as $u(R) + h_0 - \gamma x[1 - v(m)]$, now with $x = g(\omega)$ and $m = \sum_{s=1}^{S} p_s n_s y_s$. Hence, as before, we postulate that the improvement in the state of health depends on the total sum of medical expenses. Equivalently, a patient's health state improvement depends on the consumption of an aggregate good ("health care") that is efficiently produced at the lowest cost by the health care industry through the combination of medical services.

In a *fee-for-service insurance policy*, the indemnity schedule is written as $T(\omega_1, n, y)$.[16] It is an *umbrella policy* if $T(\omega_1, n, y) = I\left(\omega_1, \sum_{s=1}^{S} p_s n_s y_s\right)$, with $I(.) : \Omega_1 \times \mathbb{R}_+ \longrightarrow \mathbb{R}_+$.[17] Under the $T(.), P$ fee-for-service policy, the optimal choice of a type ω individual, denoted by $(n^{T,P}(\omega),\ y^{T,P}(\omega))$, maximizes[18]

$$u\left(w - P - \sum_{s=1}^{S} p_s n_s y_s + T(\omega_1, n, y)\right) + \gamma g(\omega) v\left(\sum_{s=1}^{S} p_s n_s y_s\right),$$

with respect to $(n, y) \in \mathcal{K}(\omega_1)$.

In this setting, a *type-contingent allocation* is written as $\{P, I_\Omega(\omega), n_\Omega(\omega), y_\Omega(\omega)\}_{|\omega\in\Omega}$ where $I_\Omega(\omega), n_\Omega(\omega)$, and $y_\Omega(\omega)$ respectively define the insurance indemnity, and the number and quantity of each medical service in state ω. The fee-for-service payment policy $T(.), P$ implements the type-contingent allocation $\{P, I_\Omega(\omega), n_\Omega(\omega), y_\Omega(\omega)\}_{|\omega\in\Omega}$ if they correspond to the same health expenses and indemnity payments in all states, that is if $\sum_{s=1}^{S} p_s n_s^{T,P}(\omega) y_s^{T,P}(\omega) = \sum_{s=1}^{S} p_s n_{\Omega s}(\omega) y_{\Omega s}(\omega)$ and $T(n^{T,P}(\omega), y^{T,P}(\omega)) =$

[16] For instance, if the policy includes copayment c_s for service $s = 1, ..., S$ and a deductible D, then $T(n, y) = \max\{\sum_{s=1}^{S} n_s(p_s y_s - c_s) - D, 0\}$.

[17] Umbrella insurance usually refers to an insurance coverage in excess of specified other policies, particularly automobile and homeowners liability policies. We use this terminology because the principle of umbrella policies is to globally protect the policyholders' wealth against uncovered risks, without reference to the specificity of these risks.

[18] For the sake of notational simplicity, we assume here that there is a unique optimal solution to the policyholder's maximization problem.

$I_\Omega(\omega)$ for all $\omega \in \Omega$. In this paper, we focus attention on umbrella policies, and Proposition 4 provides a justification to this restriction.

Proposition 4 *Any type-contingent allocation* $\{P, I_\Omega(\omega), n_\Omega(\omega), y_\Omega(\omega)\}_{|\omega\in\Omega}$ *that can be implemented by a fee-for-service payment policy* $T(.), P$*, can also be implemented by an umbrella policy* $I(.), P$*.*

Hence, in our ex post moral hazard setting, if one seeks to characterize the optimal profile of medical expenses and insurance coverage in the population with heterogeneous health states, then there is no loss of generality in restricting attention to umbrella policies. Fee-for-service policies should be chosen so as to induce this profile.[19] We have reached this conclusion by considering that the medical industry offers an aggregate good ("health care") in an efficient way. Of course, that does not mean that the optimization of fee-for-service policies is a secondary issue, but, among many reasons, this may be due to productive inefficiencies in the health industry, or because fee-for-service policies reduce transaction costs or allow insurers to better monitor the use of medical services.

[19] An umbrella policy with partial coverage at the margin and an upper limit on expenses may be induced by copayments or coinsurance on medical service, with a limit to the number and value of reimbursed services. For instance, assume that the umbrella policy can be approximated by a piecewise linear function $I(m) = (1-\alpha)\inf\{m, \overline{m}\}$, and consider the case where there is a single type of relevant act in state ω_1 (a visit to a specialist, including the purchase of drugs), and denote n, y and p, the number, quantity and price of each act. We restrict attention to stationary behaviour where the quantity is the same for each act. Then, we may define the fee-for-service policy in state ω_1 by $T(\omega_1, n, y) = (1-\alpha)p\inf\{n, \overline{n}\} \times \inf\{y, \overline{y}\}$, where $\overline{n}$ is an upper limit for the number of visits to the physician, $p\overline{y}$ is an upper expense limit per visit and α is a coinsurance rate. If $p\overline{n}\overline{y} = \overline{m}$, then in state ω_1 the policyholder chooses n and y as if she were compensated by the umbrella policy $I(m) = (1-\alpha)\inf\{m, \overline{m}\}$. Analyzing how the outcome of a more general umbrella policy $I(m)$ could be reproduced or, at least, approximated by a fee-for-service policy is an important issue in its own right, and should be the subject of further study.

With the umbrella policy, the indemnity payment $I(\omega_1, m)$ is conditioned by publicly available information ω_1. An incentive compatible mechanism now specifies the indemnity $\widehat{I}(x, \omega_1)$ and the health expenses $m(x, \omega_1)$ with $\widehat{I}(x, \omega_1) \equiv I(\omega_1, m(x, \omega_1))$. Let $f(x\,|\omega_1)$ be the density function of x conditional on ω_1, with support $[0, a(\omega_1)]$. An optimal mechanism maximizes the policyholder's expected utility

$$\int_\Omega \int_0^{a(\omega_1)} \{u(w - P + \widehat{I}(x, \omega_1) - m(x, \omega_1)) \\ + h_0 - \gamma x[1 - v(m(x, \omega_1))]\} f(x\,|\omega_1) dx d\mathbb{P}(\omega_1, \omega_2),$$

subject to the insurer's break-even constraint

$$P \geq \int_\Omega \int_0^{a(\omega_1)} \widehat{I}(x, \omega_1) f(x\,|\omega_1) dx d\mathbb{P}(\omega_1, \omega_2),$$

and to the feasibility constraints (6)-(9) where x is replaced by (x, ω_1). Conclusions would be qualitatively unchanged: given ω_1, positive indemnity are paid for all illness severity, with partial coverage at the margin and possibly bunching at the top.[20] Bunching then takes a more concrete and realistic form: insurance indemnities increase till medical expenses reach an upper limit $\overline{m}(\omega_1)$ that depends on the publicly available information on the policyholder's health state.

[20] Proofs of Proposition 1-3 and Corollaries 1-2 could be straighforwardly adapted to this more general setting. This is particularly true when the policyholder can always increase her expenses whatever the publicly available information, i.e., when $\sum_{s=1}^{S} p_s n_s y_s$ has no upper limit when $(n, y) \in \mathcal{K}(\omega_1)$. Otherwise, $m(x, \omega_1)$ is bounded by this upper limit. Thus, in addition to the information revelation mechanism, bunching could then occur at the top just because publicly available information prevents the policyholder to overspend.

4 Alternative models and robustness

4.1 Correlated background risk

The results of Section 3 extend to the case where the health level affects monetary income through an uninsurable background risk.[21] Let us assume that illness severity x randomly reduces the monetary wealth by an amount ε, with conditional c.d.f. $G(\varepsilon, x)$. We assume $G'_x(\varepsilon, x) < 0$. Thus, an increase in the illness severity level x shifts the distribution function of ε 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

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

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 straighforwardly extends Lemma 1 to this case.

Lemma 2 *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* $R(x) \equiv w - P + \widehat{I}(x) - m(x)$.

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

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 be optimal when x is uniformly distributed. Indeed, simulations show that the correlated background risk reinforces the likelihood of bunching.[22] 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 3 illustrates a case where $k = 0.01$ with bunching for the optimal contract.[23]

Figure 3

4.2 Non-separable utility

We now turn to the case where $U(R,H)$ may be non-separable between R and H.[24] It is assumed that $U(R,H)$ is increasing with respect to R and H and concave. We

[22] At first glance, one might think that bunching is less likely to be optimal when a correlated background risk creates a supplementary link between illness severity and available income, in addition to the cost of medical services. In other words, the background risk would reinforce the need for a larger insurance coverage when the severity of illness increases. In fact, this is a misleading intuition. Indeed, to preserve incentive compatibility, a \$1 increase in the insurance indemnity requires that medical expenses grow by more than \$1. Hence, if these increases are not optimal without background risk, they will be even less desirable when a correlated background risk exacerbates the variations in the expected marginal utility of wealth due to changes in medical net expenses. The simulations illustrated in Figure 3 confirm this conclusion.

[23] In Figure 3-top, indifference curves for $x = 7$ and 9 almost coincide. Figure 3-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 3 corresponds to the case of uniform distribution.

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

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 $U''_{HR} > 0$,[25] and we denote $\Phi(R,H) \equiv U'_H/U'_R$ the marginal rate of substitution between monetary wealth and health, with

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

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

Lemma 3 *Under non-separable utility, the direct revelation mechanism* $\{m(.), \widehat{I}(.)\}$ *is incentive compatible if and only if*

$$\widehat{I}'(x) = [1 - \gamma x v'(m(x))\Phi\left(R(x), H(x)\right)]\, m'(x), \tag{11}$$
$$m'(x) \geq 0, \tag{12}$$

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

$$\widehat{I}'(x) = [1 - \gamma x v'(m(x))\Phi\left(R(x), H(x)\right)]\, h(x). \tag{13}$$

[25] The assumption $U''_{HR} > 0$ is made for the sake of simplicity. One can check that Lemma 3 and following developments are still valid under more general conditions that are compatible with $U''_{HR} \leq 0$. See the proof of Lemma 2.

We have simulated the non-separable utility case with $U(R,H) = (b_0 \frac{R^{1-\alpha}}{1-\alpha} + b_1) H^{\beta}$.[26] The optimal indemnity schedule remains qualitatively similar to the characterization provided in Section 2. Figure 4-top illustrates the case of an exponential distribution with bunching.[27]

Figure 4

4.3 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 σ, which gives

$$P = (1+\sigma) \int_0^a \widehat{I}(x) f(x) dx, \tag{14}$$

instead of (2). As initially established by Arrow (1971), the optimal contract contains a straight deductible when there is a positive constant loading factor. Proposition 4 extends this characterization to the case of ex post moral hazard.

Proposition 5 *Under constant positive loading σ and with the same assumptions as Proposition 2, the optimal indemnity schedule includes a deductible $D > 0$ and an upper limit $I(\overline{m})$, that is*

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

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

[27] Figure 4-bottom adds a background risk and a loading factor, and it illustrates the optimality of a deductible, as shown in Section 4.3.

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.*

Figures 5 illustrates Corollary 4 in the case of an exponential distribution. Loading shifts the indemnity schedule rightward and creates a deductible, in addition to bunching at the top.

Figure 5

5 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 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 when $x > x^* = \inf\{x : m(x) > m^*\}$.[28] In the case of an audit, the policyholder's medical expenses are capped by the expense profile $m(x)$.[29] In other words, audit allows the insurer to monitor the policyholder's medical expenses. Thus, a type x individual chooses his 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*

[28] We here postulate that $m(x)$ is a non-decreasing function, which will be the case in what follows.

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

$I(m)$, 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 $I'(m) \leq 1$, although, as we will see, an upward discontinuity of $I(m)$ at $m = m^*$ would be optimal.[30] We denote $g(x) \equiv \widehat{I}'(x)$ when $x > x^*$, and, as previously, $h(x) = m'(x)$ for all x. For simplicity, we come back to our initial model, with separable utility and without background risk nor loading. 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

$$\widehat{I}'(x) = g(x) \text{ if } x > x^*, \tag{15}$$

$$0 \leq g(x) \leq h(x) \text{ if } x > x^*, \tag{16}$$

$$P = \left[\int_0^{x^*} \widehat{I}(x) f(x)dx + \int_{x^*}^a [\widehat{I}(x) + c] f(x)dx \right]. \tag{17}$$

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 $x > x^*$.[31] 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

[30] 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 $I'(m) \leq 1$ eliminates this reason for which an optimal solution may not exist.

[31] 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.

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

$$P \geq \int_0^{x^*} \widehat{I}(x) f(x) dx + (I^* + c)[1 - F(x^*)], \tag{18}$$

$$u'(w - P - m^* + I^*) \geq \gamma x^* v'(m^*), \tag{19}$$

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

$$\int_{x^*}^{a} \left\{ u(w - P + \widehat{I}(x) - m(x)) + h_0 - \gamma x[1 - v(m(x))] \right\} f(x) dx, \tag{20}$$

subject to (7), (15)-(17). This is a subproblem restricted to $x \in (x^*, a]$ with a non-empty set of feasible solutions.[32]

Lemma 4 *The optimal continuation allocation is such that*

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

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.*

If (18) is not binding, but the difference between its left-hand and right-hand sides is small, then increasing $\widehat{I}(x)$ over I^* is strongly constrained. Lemma 4 says that the increase in $\widehat{I}(x)$ should then be concentrated on the highest values of x, that is when

[32] $\widehat{I}(x) = I^*, m(x) = m^*, g(x) = 0, h(x) = 0$ for all $x \in (x^*, a]$ is a feasible solution because of (18). Conversely, (18) holds for any solution such that $g(x) = \widehat{I}'(x) \geq 0$ for all $x \in (x^*, a]$. Furthermore, (19) is implied by (7), (8) and (15)-(17). Thus, we may assume (18) and (19) without loss of generality.

$x > \widehat{x}$ 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 4 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.[33] Finally, it is also intuitive that 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 (20) at an optimal continuation equilibrium, where

$$A = \int_0^{x^*} \widehat{I}(x) f(x) dx. \tag{21}$$

Our global optimization problem can be rewritten as

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

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 (21). The optimal solution to this problem and the corresponding indemnity schedules are characterized as follows.

Proposition 6 *The optimal mechanism with audit is such that $x^* > 0$, with*

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

[33] 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 4 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.

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 7 *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*

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

Hence, audits 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 accompagnied by an upward jump in health expenses from $\overline{m}$ to $m(x^*)$, and insurance payment from $I(\overline{m})$ to $I(m(x^*)) = I(\overline{m}) + m(x^*) - \overline{m}$.

Proposition 6 is illustrated in Figure 6, in the exponential distribution case with $c = 0.25$. We have $x^* \simeq 4.85$. 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 6-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 $(\overline{m}, m(x^*))$.[34]

[34] The bunching of types is no more 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 $x \leq x^*$.

The dependency between the threshold x^* and the audit cost c is simulated in Figure 7. As expected, the larger the audit cost, the larger the threshold above which it is optimal to audit health care expenses.

Figures 6 and 7

Finally, for the sake of realism, we may enrich the previous characterization by assuming that the the insurance premium is computed with a loading factor $\sigma > 0$, which gives[35]

$$P = (1+\sigma)\left[\int_0^{x^*} \widehat{I}(x)f(x)dx + \int_{x^*}^{a} [\widehat{I}(x)+c]f(x)dx\right].$$

Unsurprisingly, the optimal indemnity schedule, under the continuity assumption for $I(m)$, now mixes a deductible $D > 0$ and a maximum of out-of-pocket expenses $\overline{m} - I(\overline{m})$.

Proposition 8 *Under constant positive loading σ, the optimal indemnity schedule includes a deductible $D > 0$. Under the same assumptions as Proposition 2, we have*

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

Proposition 8 is illustrated in Figure 8, here also in the exponential distribution case. The larger the loading factor σ, the larger the deductible D.

Figure 8

[35] The fact that the loading also applies to audit costs is an innocuous assumption. It amounts to replacing c with $(1+\sigma)c$.

6 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. In this paper, we have characterized an optimal health care reimbursement scheme under various assumptions. We have started with the most simple case of separable utility. Not in full generality, but under reasonable assumptions, 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. Coinsurance is indeed the most usual way to mitigate health expenses under ex post moral hazard, but upper limits on coverage also frequently exist through caps on various types of health expenditures.[36] Insurance contracts under ex post moral hazard trade off risk sharing and incentives, but bunching high-severity low-probability illnesses is also likely part of the optimal solution to this trade-off. However, the optimal indemnity schedule, and thus its slope and upper limit, depend on the publicly available information on the policyholder's health state. We have also shown that this characterization remains valid in the case of a correlated background risk and when utility is non-separable between wealth and health.

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 insurance loading, 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

[36] See Cutler and Zeckhauser (2000).

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 upper limit for out-of-pocket expenses.

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 limits to out-of-pocket expenses. This complexity reflects what we frequently observe in the real world, all these ingredients being mixed, with more complete coverage and limits to out-of-pocket expenses, for 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.[37]

Although we think that it is of utmost importance to establish such general principles for the design of health insurance, it remains true that translating these principles into concrete strategy is crucial. This concern is related to the design of fee-for-service reimbursement rules that best fit the characteristics of an optimal indemnity schedule, and that may allow insurers to better monitor the use of medical services. It is also related to the role of health care providers. We have restricted ourselves to the most simple case where doctors and hospitals act as "perfect agents" of policyholders, without any other behavior rule than choosing the level of health expenses that is in the policyholders' best interests. The development of various forms of managed care in many countries reflects the fact that the objectives and constraints of health care providers may affect the way patients use medical services. This is a dimension that insurers and policy makers always keep in mind when they design health insurance

[37] 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.

plans.

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, 2001.

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.

Cutler, D.M. and R.J. Zeckhauser, 2000, "The anatomy of health insurance", in *Handbook of Health Economics*, A.J. Culyer and J.P. Newhouse (eds.), Vol.1, Chapter 11, 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.

Evans, W.N. and W.K. Viscusi, 1991, "Estimation of state dependent utility func-

tions 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.

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

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

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

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, N°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.

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

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 con-

straint are respectively rewritten as

$$A(x) + \gamma x B(x) \geq A(\widetilde{x}) + \gamma x B(\widetilde{x}) \text{ for all } x, \widetilde{x}, \tag{22}$$

$$w \geq \int_0^a [u^{-1}(A(x)) + v^{-1}(B(x)]f(x)dx, \tag{23}$$

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 (22),(23) 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

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

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 nondecreasing. Since $I(.)$ is nondecreasing, $\widehat{I}(.) \equiv I(m(.))$ is also nondecreasing.

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[38] 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

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

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 nondecreasing, 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 $\widehat{I}_1(x) \equiv \widehat{I}_0(x)$.

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

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

$$\frac{\partial^2 V(x,\widetilde{x})}{\partial \widetilde{x}^2} \mid_{\widetilde{x}=x} \leq 0, \tag{25}$$

at any point of differentiability. (24) and (25) 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 (24) 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 (25) 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 λ and $\delta(x)$ be Lagrange multipliers respectively for (2) and (9). The Hamiltonian is written as

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

The optimality conditions are

$$\varphi(x) \equiv \mu_1(x)\left[1 - \frac{\gamma x v'(m(x))}{u'(R(x))}\right] + \mu_2(x) \leq 0, = 0 \text{ if } h(x) > 0, \tag{26}$$

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

$$\mu_2'(x) = [u'(R(x)) - \gamma x v'(m(x))]f(x)$$
$$+\mu_1(x)h(x)\gamma x\left[\frac{v''(m(x))u'(R(x)) + v'(m(x))u''(R(x))}{u'(R(x))^2}\right], \tag{28}$$

$$\mu_1(a) = \mu_2(a) = 0, \tag{29}$$

$$\lambda - \int_0^a \left[u'(R(x))f(x) + \mu_1(x)h(x)\gamma x\frac{v'(m(x))u''(R(x))}{u'(R(x))^2}\right]dx = 0, \tag{30}$$

with $\delta(x) \geq 0$ and $\delta(x) = 0$ if $\widehat{I}(x) > 0$. A tedious but straightforward calculation using (27) and (28) leads to

$$\varphi'(x) = [\lambda f(x) - \delta(x)]\left[1 - \frac{\gamma x v'(m(x))}{u'(R(x))}\right] - \gamma\mu_1(x)\frac{v'(m(x))}{u'(R(x))}. \tag{31}$$

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

$$\widehat{m}(x) \equiv \arg\max_{\widetilde{m} \geq 0}\{u(w - P - \widetilde{m}) + \gamma x v(\widetilde{m})\}, \tag{32}$$

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

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

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 $0 < \varepsilon < x_0$.[39]

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 $x \in (x_0 - \varepsilon, x_0]$. (31) 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 ε 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 (31) then gives $\mu_1(x_0)_+ = 0$. Thus, $\mu_1(x)$ is continuous at x_0.

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

Integrating $\mu_1'(x)$ given by (27) and using (29) and (30) 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 $\delta(x_0 + \eta_0) = 0$.

[39] 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 $x \in (x_0, x_0 + \varepsilon)$.

Hence (27) 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 (27) 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 (31) 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 $[0, x_0]$.[40] (27) 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.

[40] Note that (27) and $\mu_1(x) = \mu_1'(x) = 0$ for all $x \in [0, x_0]$ imply that $\delta(x)$ is continuous in this interval.

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 9 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 9

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 (27), (31) and $\delta(x) = h(x) = 0$ if $x \in [x_1, x_2]$ yields

$$\varphi'(x) = \lambda[1 - \frac{\gamma x v'(m_0)}{u'(R_0)}]f(x) - \gamma\mu_1(x)\frac{v'(m_0)}{u'(R_0)}, \tag{33}$$

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

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

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 $\varphi'(x_2)_- \geq 0$. (31), $\delta(x_2) = 0$ and $\mu_1(x_2) > 0$[41] give $\gamma x_2 v'(m_0) \leq u'(R_0)$. If $f(x)$ is non-increasing and

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

condition (10) holds, then we have $\Lambda'(x) \geq 0$ if $x \leq x_2$.[42] 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 (31) 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 (31) 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 (31) and $h(x) = \delta(x) = 0$ if $x \in [\overline{x}, 1]$ gives

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

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

[42] Assume that $f(x)$ is non-increasing. Let $x \leq x_2$. If $d^2 \ln f(x)/dx^2 \geq 0$, then using $\gamma x_2 v'(m_0) \leq u'(R_0)$ directly implies $\Lambda'(x) \geq 0$. If $d^2 \ln f(x)/dx^2 < 0$, then using (10) and $\gamma v'(m_0)/u'(R_0) < x \leq x_2$ also yields $\Lambda'(x) \geq 0$.

Assume $f(a) = f'(a) = 0$ and $f''(a) > 0$. Suppose $\overline{x} = a$ and thus $h(x) > 0$ for all $x \in [0, a]$.[43] We also have $\varphi'(x) = \delta(x) = 0$ for all x. Differentiating (31) 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 $J(a) = J'(a) = J''(a) = 0$ and $h(a) = 0$.*

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

$$J'(x) = -\frac{d \ln f(x)}{dx}\mu_1'(x) - \frac{d^2 \ln f(x)}{dx^2}\mu_1(x)$$
$$+ f'(x)[2\lambda - u'(R(x))] - f(x)u''(R(x))R'(x). \tag{34}$$

Using $\mu_1(a) = f(a) = 0, \delta(x) = 0$ for all x and (27) gives $\mu_1'(a) = 0$. (34) 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 $J''(a) \geq 0$.

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

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 (27) and $\varphi'(x) \equiv 0$ by using the L'Hôpital's rule twice. Furthermore, (27) gives $\mu_1''(a) = 0$ and (34) then yields $J''(a) = f''(a)[2\lambda - u'(R(a))]$, which implies $u'(R(a)) = 2\lambda$.

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

Step 3: *Let* $\xi(x) \equiv u'(R(x))\varphi'(x)$*, where* $\varphi(x)$ *is defined by (26). 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 $\xi(x) = \lambda f(x)\Delta_1(x) - \gamma\Delta_1(x)$, 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 (27) 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), (31) 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

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

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 $t \in \mathbb{R}$:

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

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 Proposition 4

Let $I_\Omega(.), n_\Omega(.), y_\Omega(.), P$ be a type-contingent allocation that is implemented by the fee-for-service policy $T(.), P$. Let $I(.) : \Omega_1 \times \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ be defined by

$$\begin{aligned} &I(\omega_1, m) = I_\Omega(\omega) \text{ if there exists } \omega = (\omega_1, \omega_2) \in \Omega \\ &\text{such that } \sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega) y_{\Omega s}(\omega) = m, \\ &I(m) = 0 \text{ otherwise.} \end{aligned}$$

We may first check that $I(.)$ is well-defined. Indeed, assume that there exist $\omega = (\omega_1, \omega_2) \in \Omega$ and $\omega' = (\omega_1, \omega_2') \in \Omega$ such that

$$\begin{aligned} &\sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega) y_{\Omega s}(\omega) = \sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega') y_{\Omega s}(\omega') = m, \\ &I_\Omega(\omega') > I_\Omega(\omega). \end{aligned}$$

Since the allocation is implemented by $T(.), P$, we have $T(\omega_1, n_{\Omega s}(\omega'), y_{\Omega s}(\omega')) = I_\Omega(\omega') > I_\Omega(\omega) = T(\omega_1, n_{\Omega s}(\omega), y_{\Omega s}(\omega))$. Thus, under the fee-for-service policy $T(.), P$, type ω' policyholders would be better off by choosing $(n_\Omega(\omega'), y_\Omega(\omega'))$ instead of $n^{T,P}(\omega'), y^{T,P}(\omega')$, since $(n_\Omega(\omega), y_\Omega(\omega))$ is a feasible choice (i.e., it is in $\mathcal{K}(\omega_1)$) with the same cost (and thus with the same improvement in health) and with a larger insurance indemnity. This would contradict the definition of $n^{T,P}(.), y^{T,P}(.)$.

Secondly, let us show that $I_\Omega(.), n_\Omega(.), y_\Omega(.), P$ is implemented by $I(.), P$.

Let $\omega = (\omega_1, \omega_2) \in \Omega$ and let $m^0 = \sum_{s=1}^{S} p_s n_s^0 y_s^0 \geq 0$ with $(n^0, y^0) \in \mathcal{K}(\omega_1)$. Assume first that there exists $\omega' = (\omega_1, \omega_2') \in \Omega$ such that $m^0 = \sum_{s=1}^{S} p_s n_{\Omega s}(\omega') y_{\Omega s}(\omega')$. From the definition of $I(.)$ and $n^{T,P}(.), y^{T,P}(.)$, and using the fact that $P, T(.)$ implements

$P, I_\Omega(.), n_\Omega(.), y_\Omega(.)$, we may write

$$
\begin{aligned}
& u(w - P - m^0 + I(\omega_1, m^0)) + \gamma g(\omega) v\left(m^0\right) \\
& = u(w - P - \sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega') y_{\Omega s}(\omega') + I_\Omega(\omega')) \\
& + \gamma g(\omega) v\left(\sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega') y_{\Omega s}(\omega')\right) && (35) \\
& = u\left(w - P - \sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega') y_{\Omega s}(\omega') + T(\omega_1, n_{\Omega s}(\omega'), y_{\Omega s}(\omega'))\right) \\
& + \gamma g(\omega) v\left(\sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega') y_{\Omega s}(\omega')\right) && (36) \\
& \leq u\left(w - P - \sum\nolimits_{s=1}^{S} p_s n_s^{T,P}(\omega) y_s^{T,P}(\omega) + T(\omega_1, n^{T,P}(\omega), y^{T,P}(\omega))\right) \\
& + \gamma g(\omega) v\left(\sum\nolimits_{s=1}^{S} p_s n_s^{T,P}(\omega) y_s^{T,P}(\omega)\right) && (37) \\
& = u\left(w - P - \sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega) y_{\Omega s}(\omega) + I\left(\omega_1, \sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega) y_{\Omega s}(\omega)\right)\right) \\
& + \gamma g(\omega) v\left(\sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega) y_{\Omega s}(\omega)\right), && (38)
\end{aligned}
$$

where: (35) follows from the definition of $I(.)$ and the assumption made about m^0, (36) holds because $P, T(.)$ implements $P, I_\Omega(.), n_\Omega(.), y_\Omega(.)$, (37) follows from the definition of $n_s^{T,P}(.), y_s^{T,P}(.)$ and (38) holds because of the definition of $I(.)$ and $P, T(.)$ implements $P, I_\Omega(.), n_\Omega(.), y_\Omega(.)$.

Assume now that there does not exist $\omega' = (\omega_1, \omega_2') \in \Omega$ such that $m^0 = \sum_{s=1}^{S} p_s n_{\Omega s}(\omega') y_{\Omega s}(\omega')$. Then we may write

$$
\begin{aligned}
& u(w - P - m^0 + I(\omega_1, m^0)) + \gamma g(\omega) v\left(m^0\right) \\
& = u(w - P - \sum\nolimits_{s=1}^{S} p_s n_s^0 y_s^0) + \gamma g(\omega) v\left(\sum\nolimits_{s=1}^{S} p_s n_s^0 y_s^0\right) && (39) \\
& \leq u\left(w - P - \sum\nolimits_{s=1}^{S} p_s n_s^0 y_s^0 + T(\omega_1, n_s^0, y_s^0)\right) + \gamma g(\omega) v\left(\sum\nolimits_{s=1}^{S} p_s n_s^0 y_s^0\right) && (40) \\
& \leq u\left(w - P - \sum\nolimits_{s=1}^{S} p_s n_s^{T,P}(\omega) y_s^{T,P}(\omega) + T(\omega_1, n^{T,P}(\omega), y^{T,P}(\omega))\right) \\
& + \gamma g(\omega) v\left(\sum\nolimits_{s=1}^{S} p_s n_s^{T,P}(\omega) y_s^{T,P}(\omega)\right) && (41) \\
& = u\left(w - P - \sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega) y_{\Omega s}(\omega) + I\left(\omega_1, \sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega) y_{\Omega s}(\omega)\right)\right) \\
& + \gamma g(\omega) v\left(\sum\nolimits_{s=1}^{S} p_s n_{\Omega s}(\omega) y_{\Omega s}(\omega)\right), && (42)
\end{aligned}
$$

where (39) follows from the definition of $I(.)$, (40) results from $T(.) \geq 0$, (41) follows from the definition of $n_s^{T,P}(.), y_s^{T,P}(.)$, and (42) holds because of the definition of $I(.)$ and $P, T(.)$ implements $P, I_\Omega(.), n_\Omega(.), y_\Omega(.)$.

Thus, in both cases, $(n_\Omega(\omega), y_{\Omega s}(\omega))$ is an optimal choice of type ω policyholders with the above-defined umbrella policy, and with the same indemnity as with the fee-for-service policy $T(.), P$, with $I(m_\Omega(\omega)) = I_\Omega(\omega)$, which shows that $P, I(.)$ implements $P, I_\Omega(.), m_\Omega(.)$.

Proof of Lemma 2

Similar to Lemma 1, with straightforward adaptation.

Proof of Lemma 3

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 (11) is a necessary condition for incentive compatibility. (11) 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[44]

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

[44] On can check that $A'_{\widetilde{x}}(x, \widetilde{x})_{|\widetilde{x}=x} > 0$ 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 $U''_{RH} < 0$. Thus Lemma 3 is valid under more general conditions than the ones we have considered in Section 4.

Thus incentive compatibility gives (12). Conversely, assume that (11) and (12) 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 5

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 (31) by

$$\varphi'(x) = [\lambda(1+\sigma)f(x) - \delta(x)]\left[1 - \frac{\gamma x v'(m(x))}{u'(R(x))}\right] - \gamma\mu_1(x)\frac{v'(m(x))}{u'(R(x))}. \tag{43}$$

and λ by $\lambda(1+\sigma)$ in (27).

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

$$h(x) > 0, \tag{44}$$
$$1 - \frac{\gamma x v'(m(x))}{u'(R(x))} > 0, \tag{45}$$

if $x < \varepsilon$. (44) implies $\varphi(x) = \varphi'(x) = 0$ if $x < \varepsilon$. Furthermore, using (27) (in which λ is replaced by $\lambda(1+\sigma)$), (30) 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. (43) and (45) then yields $\varphi'(x) > 0$, 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 $D = m(d)$.[45]

Figure 10

Proof of Lemma 4

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

$$\int_{x^*}^{a} \Big\{u(w - P + \widehat{I}(x) - m(x)) + h_0 - \gamma x[1 - v(m(x))]\Big\} f(x)dx, \tag{46}$$

subject to (7) and (15)-(17).

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 λ be Lagrange multipliers respectively for (16) and (17) in this subproblem.[46] The Hamiltonian is written as

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

and the optimality conditions are

$$\mu_1(x) - \eta(x) \leq 0, = 0 \text{ if } g(x) > 0, \tag{47}$$

$$\mu_2(x) + \eta(x) = 0, \tag{48}$$

$$\mu_1'(x) = [\lambda - u'(R(x))]f(x), \tag{49}$$

$$\mu_2'(x) = [u'(R(x)) - \gamma x v'(m(x))]f(x)\ , \tag{50}$$

[45] Note however, that we may have $I'(D_+) = 0$.

[46] We can straighforwardly check that (8) is not binding in this subproblem.

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 $x \in \mathcal{V}$. (47) gives $\mu_1(x) = 0$, and thus $\mu_1'(x) = 0$ for all $x \in \mathcal{V}$. Then (49) 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 $x \in \mathcal{V}$. (47) and (48) yield $\mu_1(x) = -\mu_2(x) = \eta(x)$, and thus $\mu_1'(x) = -\mu_2'(x)$, for all $x \in \mathcal{V}$. (49) and (50) 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 6

Let $\mu_1(x)$ and $\mu_2(x)$ be costate variables respectively for $\widehat{I}(x)$ and $m(x)$ and let $\delta(x)$ and λ be Lagrange multipliers respectively for (9) and (21). The Hamiltonian is written as in the proof of Proposition 1, and the optimality conditions (26), (27) and (28) 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

$$V_1' - \mu_2(x^*) = 0, \tag{51}$$

$$V_2' - \mu_1(x^*) = 0, \tag{52}$$

$$V_3' + \{u(R^*) + h_0 - \gamma x^*[1 - v(m^*)]\}f(x^*)$$
$$-\mu_1(x^*)\frac{\gamma x^* v'(m^*)}{u'(R^*)} - [\lambda - \delta(x^*)]I^* \leq 0, = 0 \text{ if } x^* > 0, \tag{53}$$

$$V_4' - \int_0^{x^*} \left[u'(R(x))f(x) + \mu_1(x)h(x)\gamma x\frac{v'(m(x))u''(R(x))}{u'(R(x))^2}\right] dx = 0, \tag{54}$$

$$V_5' + \lambda = 0, \tag{55}$$

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 (28) 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 (53) 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 (53) 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*

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

Identical to the proof of Proposition 2.

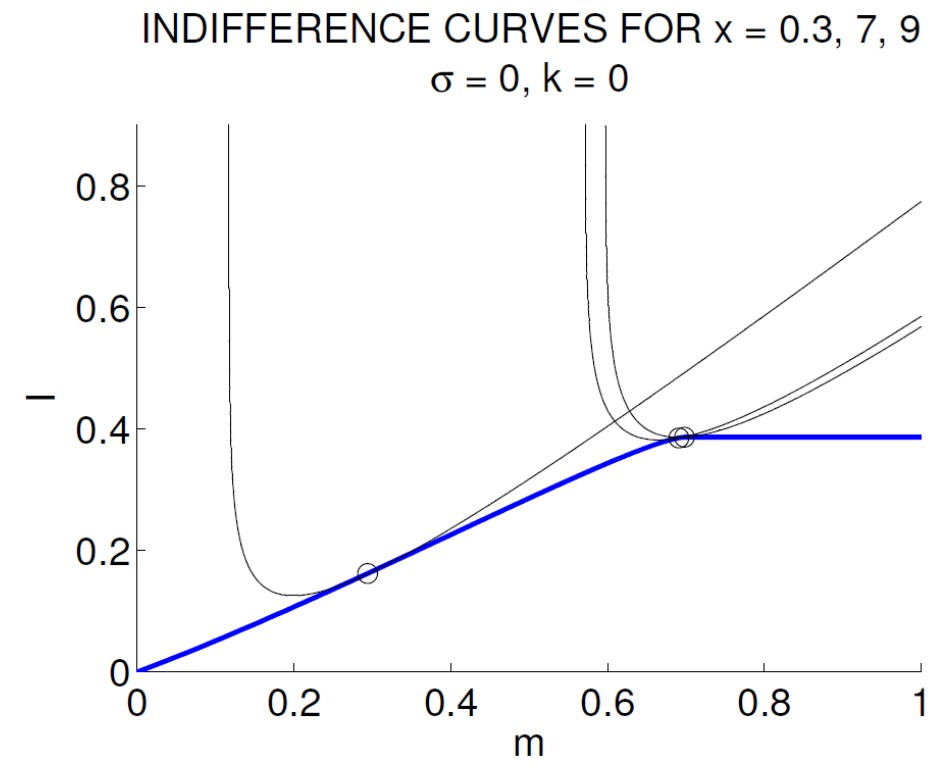
Finally, $\mu_1(x^*) > 0$ shows that there is an upward discontinuity in $m(x)$ and $\widehat{I}(x)$ at $x = x^*$.

Proof of Proposition 7

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

Proof of Proposition 8

The Proposition follows from a straightforward adaptation of the proof of Proposition 5.



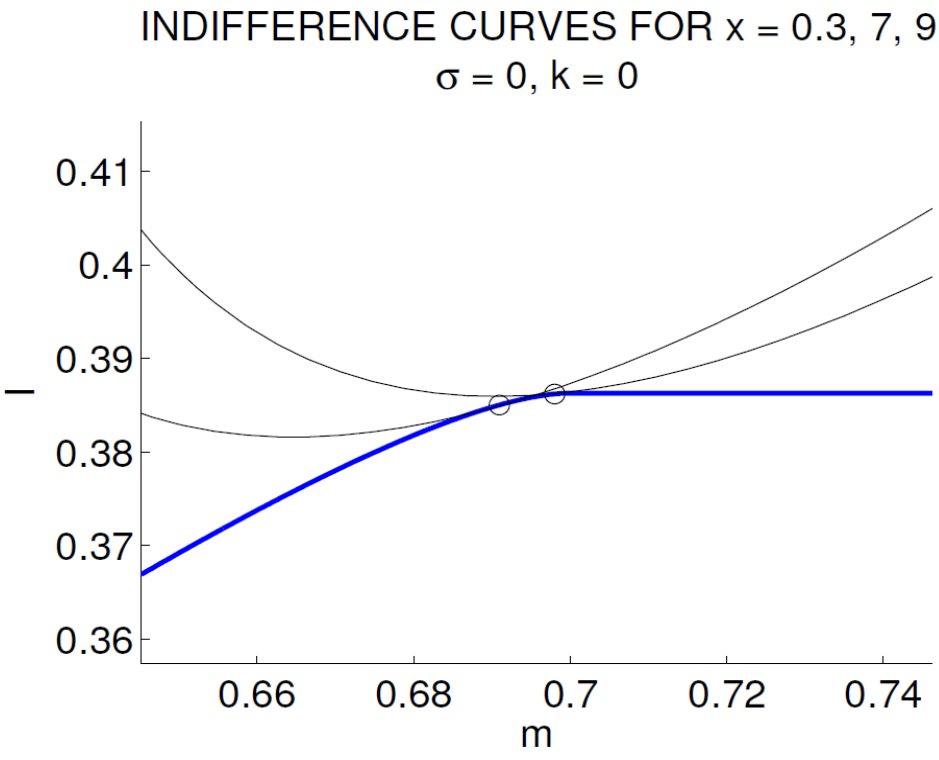
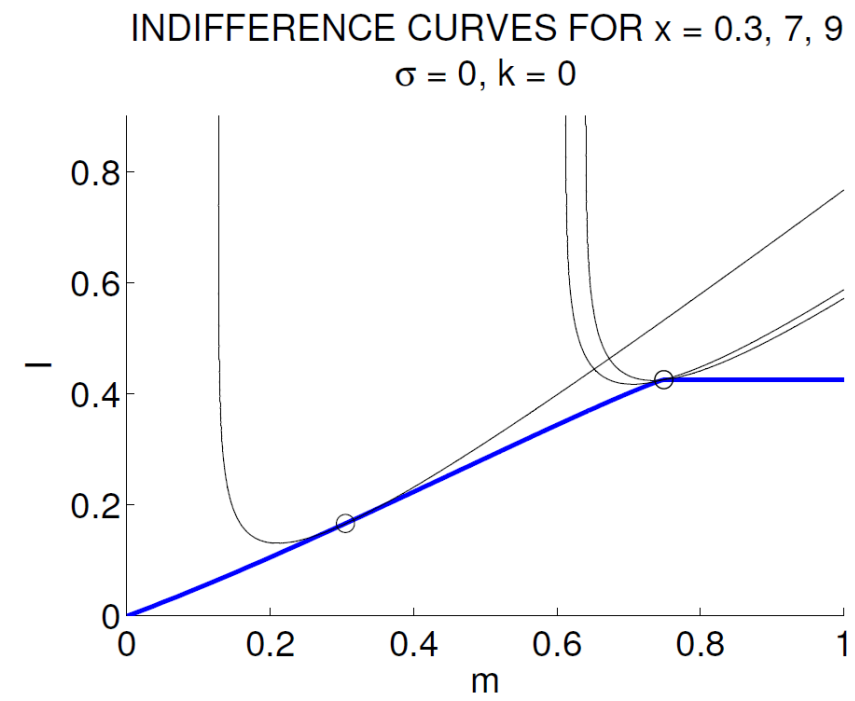


Figure 1

Uniform distribution – No bunching



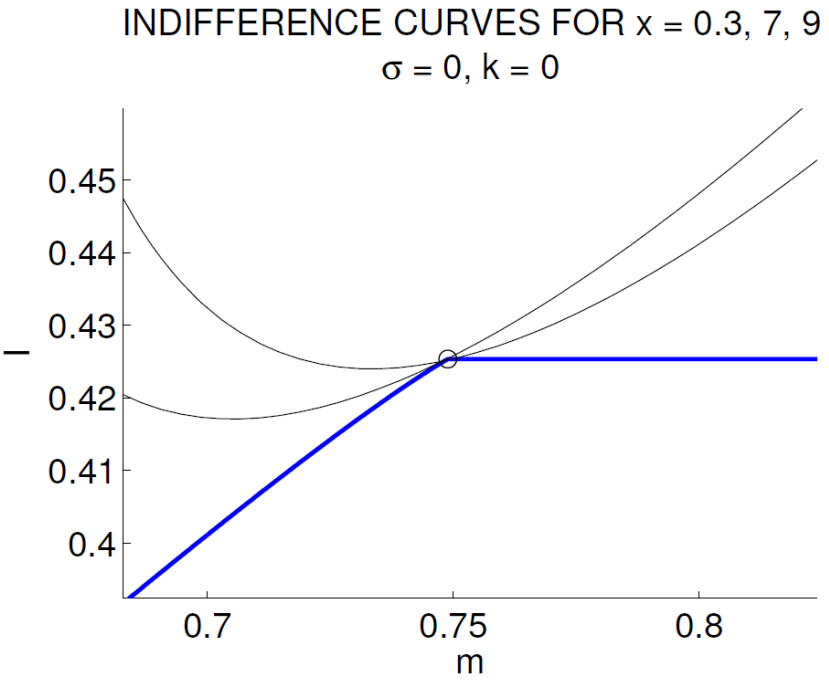
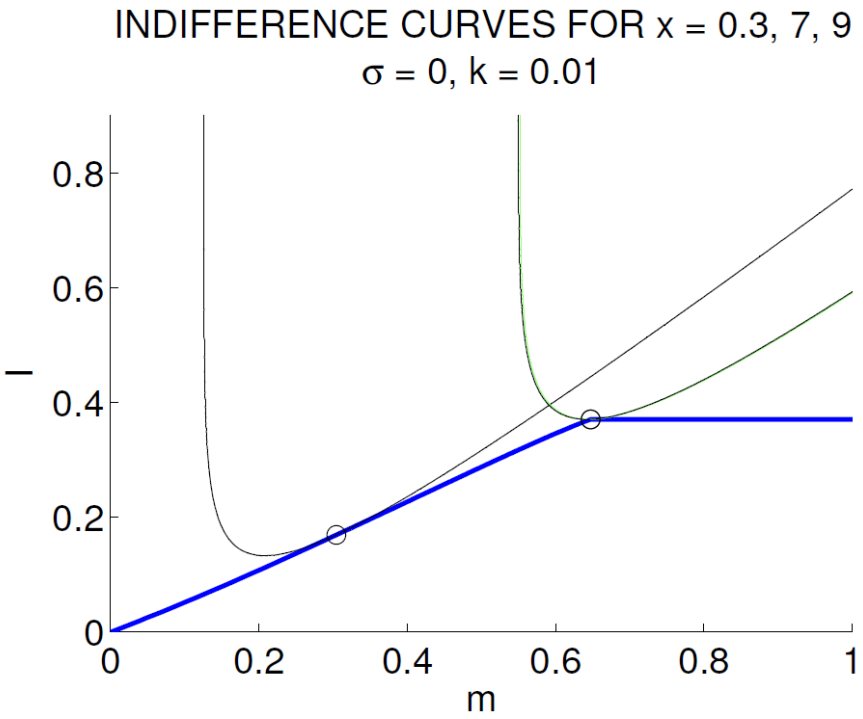


Figure 2

Exponential distribution - Bunching



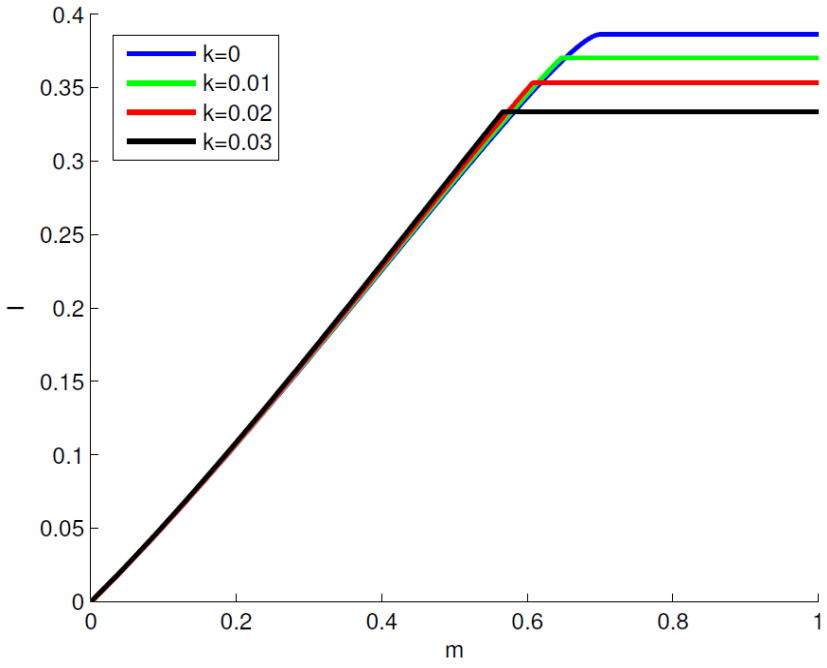
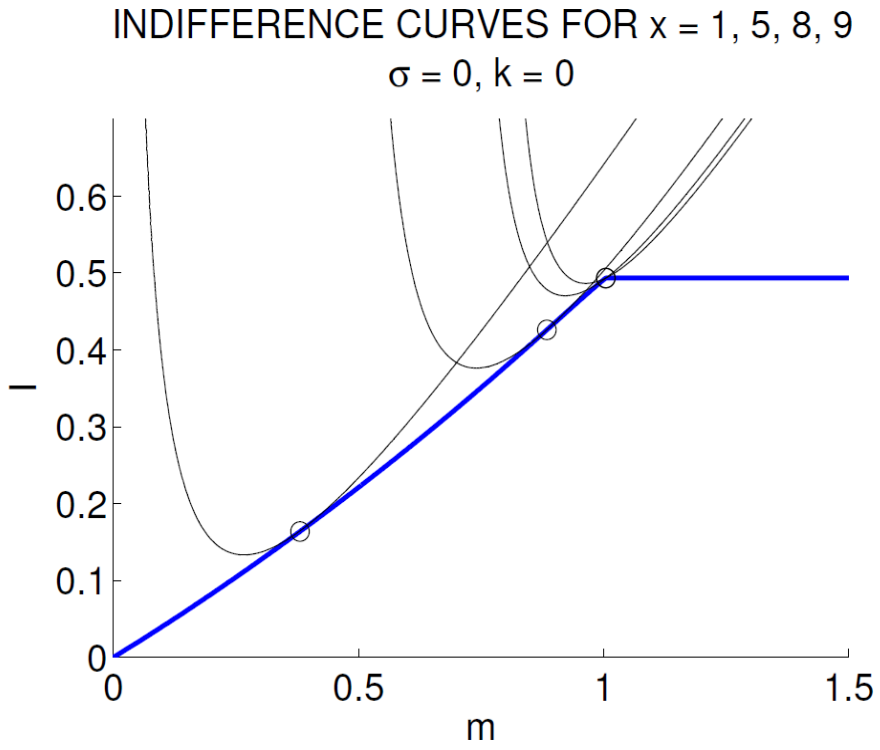


Figure 3

Uniform distribution

Case where the background risk creates bunching



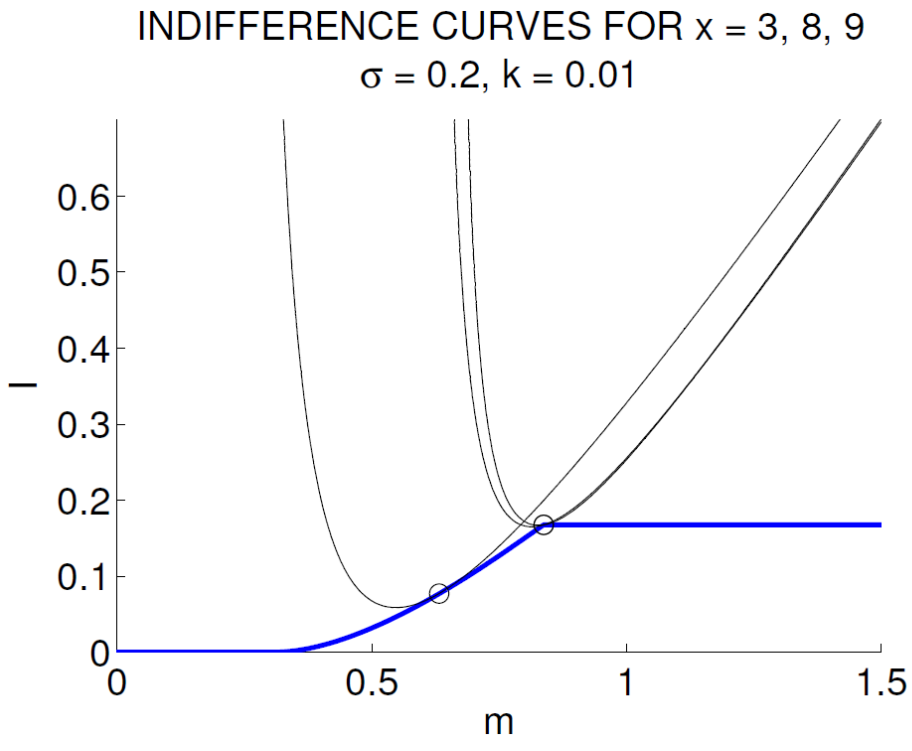
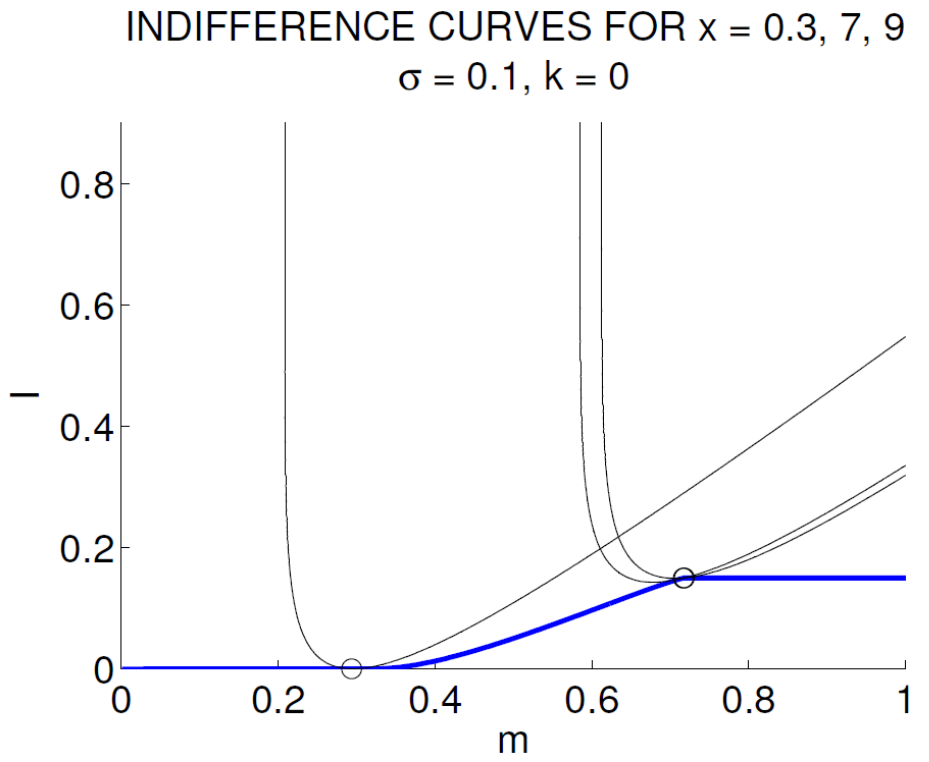


Figure 4

Non-separable utility – Exponential distribution



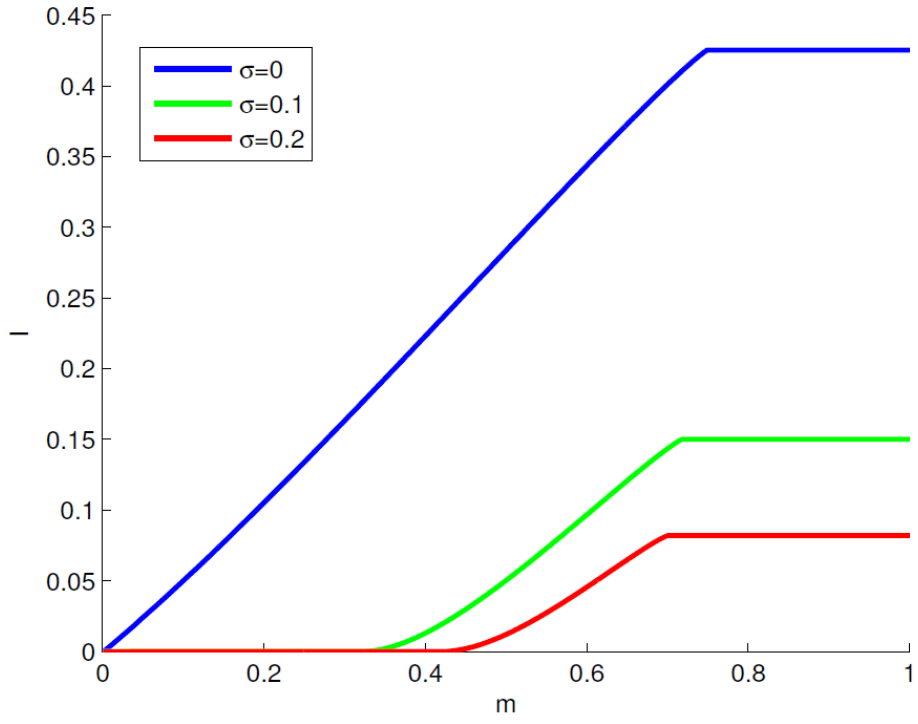
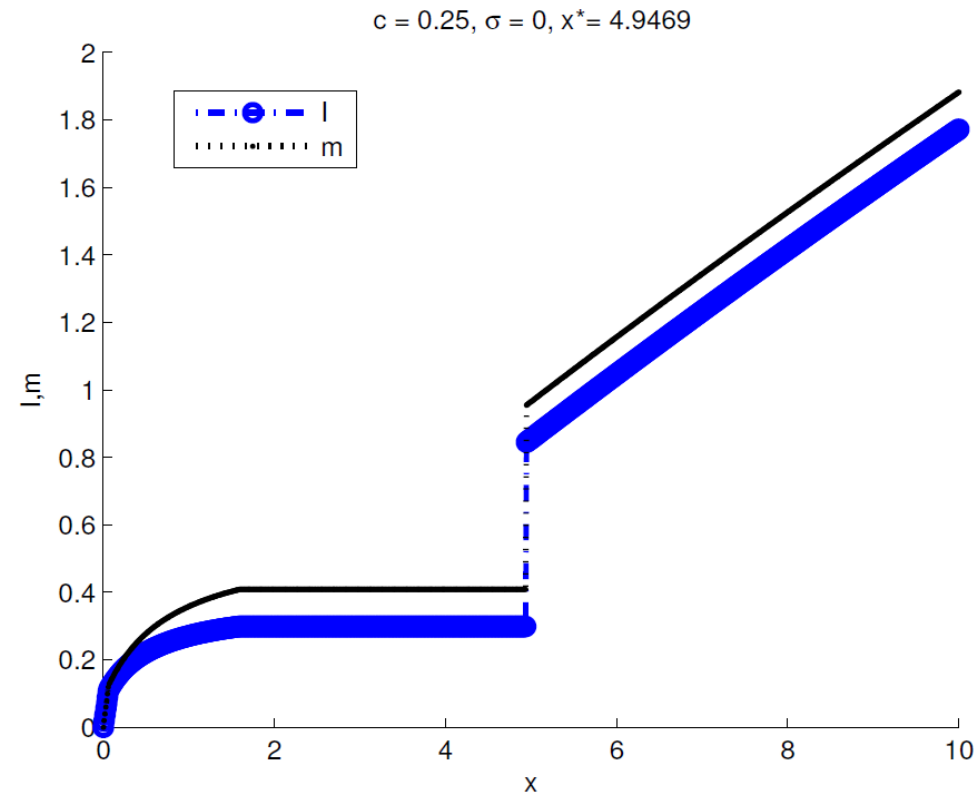


Figure 5

Exponential distribution - A deductible is optimal under loading



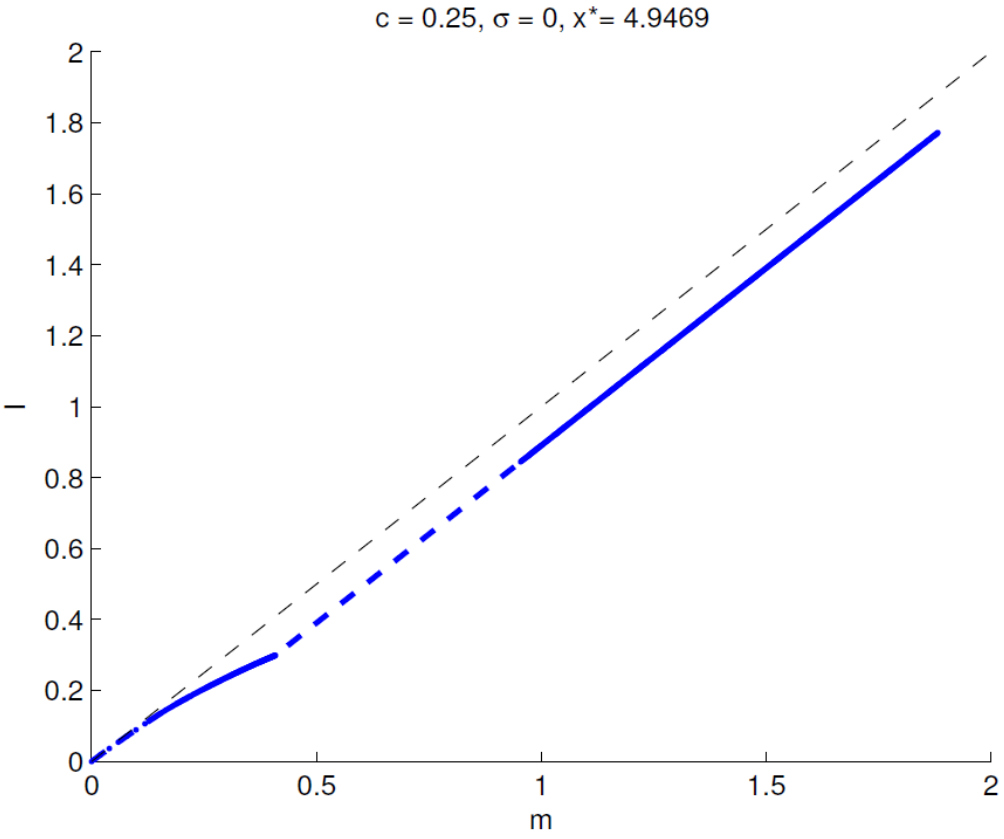


Figure 6

Exponential distribution: Auditing without loading

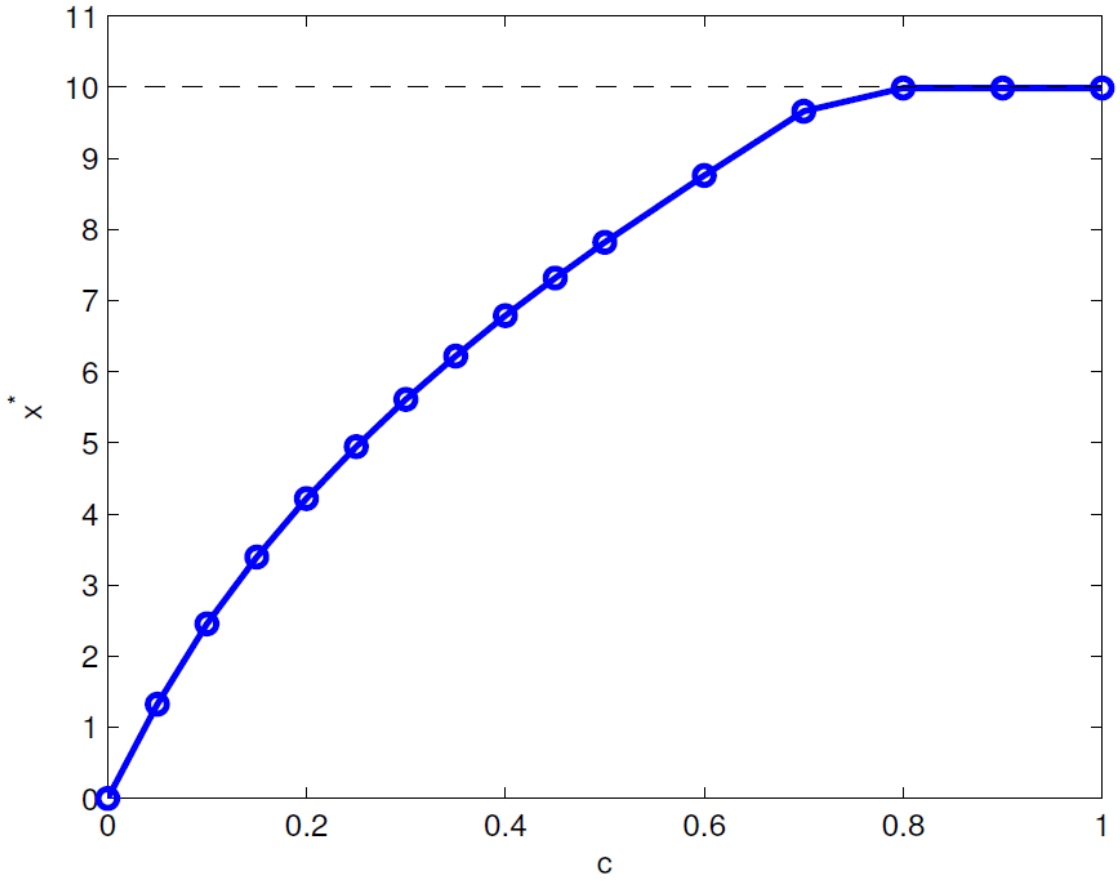


Figure 7

Dependency between threshold x^* and audit cost c

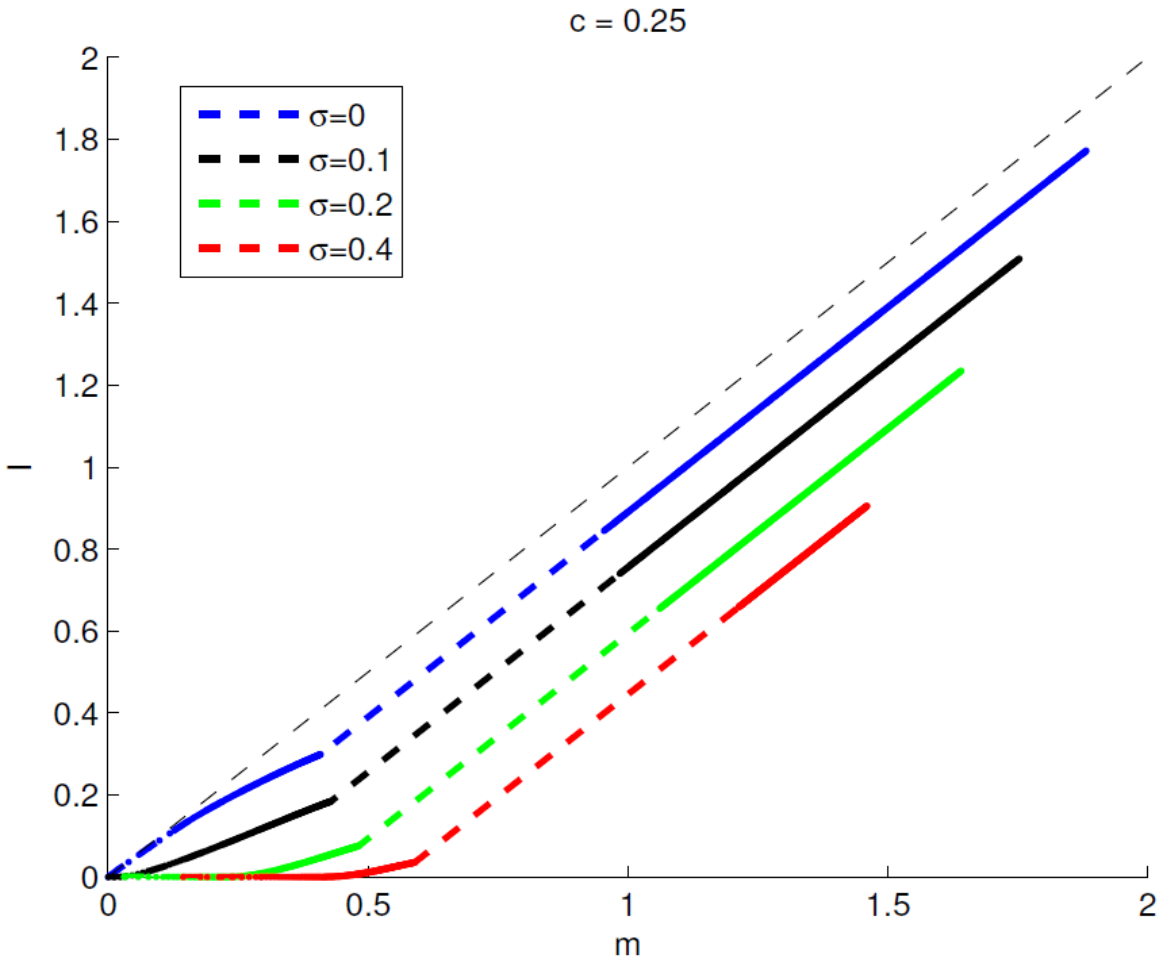


Figure 8

Exponential distribution - Loading and auditing

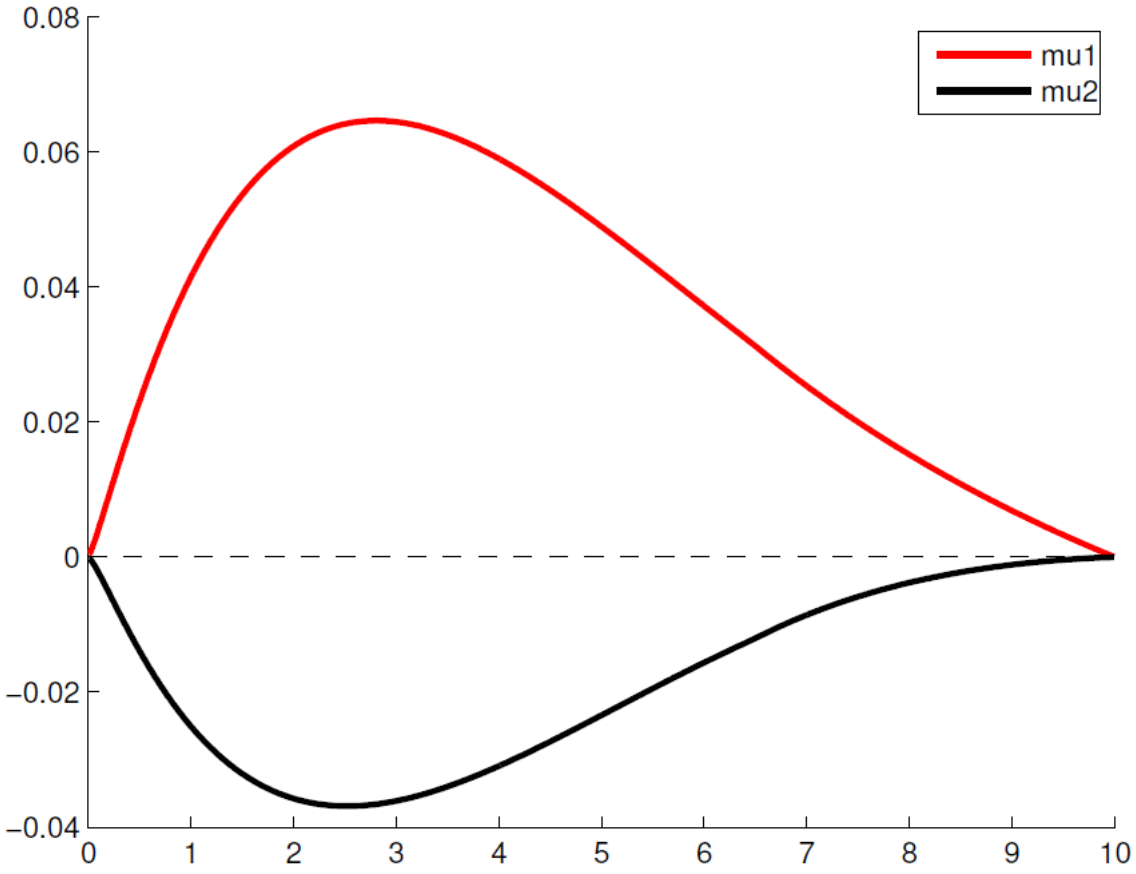


Figure 9

Trajectories of co-state variables

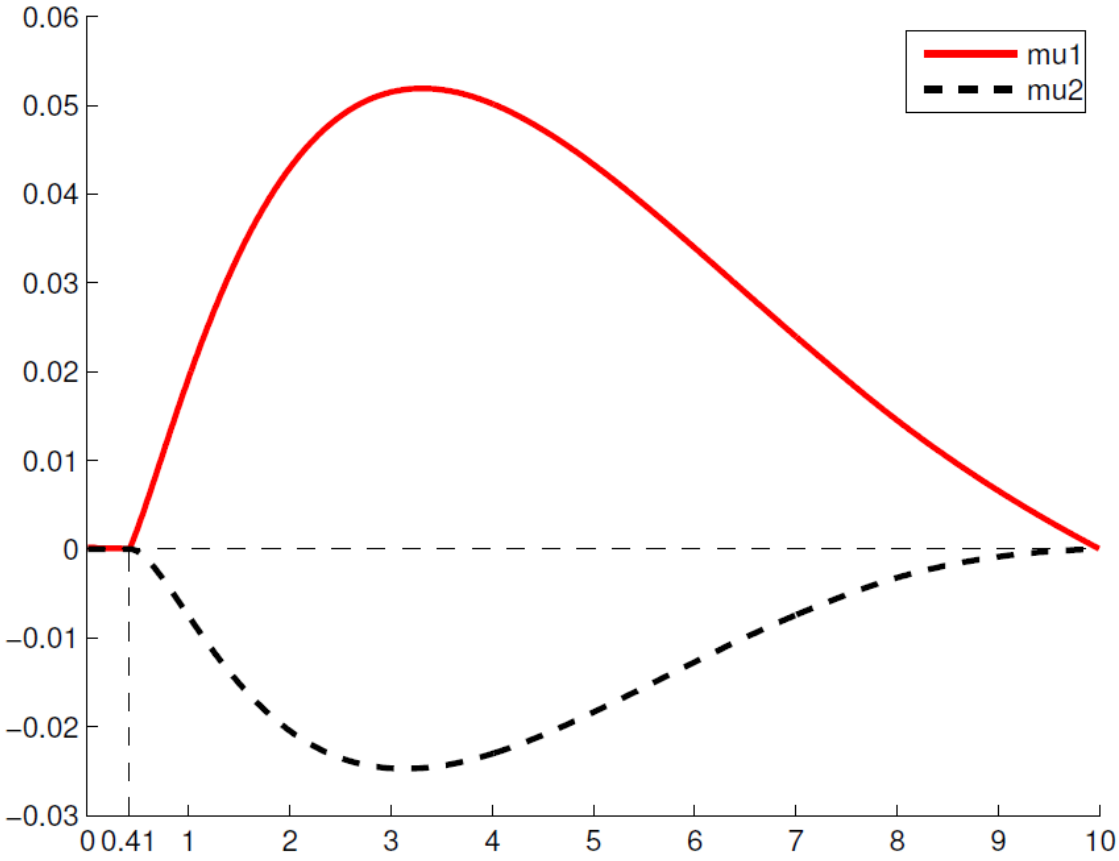


Figure 10

Co-state variables under loading