Medical Savings Accounts or Deductible

by Ronny Klein
Medical Savings Accounts or Deductible.
A Theoretical and Simulation-based Comparison.

Ronny Klein∗

April 2004

Abstract
A medical savings account (MSA) is a health insurance which includes a savings account: if ill, the insured firstly spends the money out of the account until it is depleted. Further costs are paid out-of-pocket and by the insurer. This paper compares a MSA and a common deductible insurance (DI) using a two-period expected utility model. Furthermore, in a simulation of different MSA plans based on a real world sample of insurance claims, we search for the policy which maximizes an individual’s expected utility as well as the average welfare of the entire risk pool. The results suggest that saving unused co-payments to pay medical costs in older years cannot only increase the overall welfare but also reduce the total medical expenditures. Compared to a rebate policy, a MSA has the advantage that it addresses directly the age-related increase of medical costs.

JEL classification: D80, D91, G22, I10
Keywords: Insurance, Medical Savings Account, Deductible, Moral Hazard

1 Introduction
Individual savings accounts as a supplement to a social security insurance are widely accepted within the context of pension funding systems as a way to temper the financial burden of the aging society. The introduction of the so called Riester Rente in Germany in the year 2002 is one of the recent examples for the popularity of this idea. In 2002, Chile introduced individual unemployment accounts which are complemented by a public fund in the

∗I acknowledge the support by the Swiss insurance company CSS which has allowed me to use their comprehensive dataset for my analysis. Furthermore, I’m grateful to Stefan Felder and Andreas Werblow who have helped me to obtain the sample. Contact: Martin-Luther-University Halle-Wittenberg, Department of Economics, 06099 Halle, Phone: +49/345/5523443, Fax: +49/345/5527127, Email: r.klein@wiwi.uni-halle.de.
case of depletion.\footnote{See Feldstein and Altman (1998) and Orszag and Snower (2002) for an analysis of unemployment accounts.} But, also in health care financing, there is growing interest in introducing individual health accounts in order to create incentives to utilize health care more prudently and to encourage disease prevention activities.

In 1996 the 104th Congress released the \textit{Health Insurance Portability and Accountability Act}\footnote{HIPAA, Public Law 104-191, August 21, 1996.} and established a pilot program for medical savings accounts (MSA) in the USA. The Internal Revenue Service (2001) defines a MSA as follows: \footnote{An Archer MSA is a tax-exempt trust or custodial account with a financial institution (like a bank or an insurance company) in which you can save money for future medical expenses. This account must be used in conjunction with an HDHP.} “An Archer MSA is a tax-exempt trust or custodial account with a financial institution (like a bank or an insurance company) in which you can save money for future medical expenses. This account must be used in conjunction with an HDHP.”\footnote{HDHP: high deductible health plan. High deductible means higher than with typical health plans.} In particular, the tax exemption of contributions to a MSA should be an incentive to sign such a contract. Bunce (2001) gives an overview of the development of MSAs in the USA. Although during the pilot program the public interest was surprisingly low, according to the concluding report of the GAO (1998), the program has been extended until the end of 2002.

In Singapore, the so called \textit{Medisave accounts} were already established nationwide in 1984. Massaro and Wong (1995, 1996) argue that this is the reason for the relative low health expenditures of 3.1\% of the GNP in comparison with other Asian countries and all the more with western industrial nations.

According to Matisonn (2000), MSAs became a popular health insurance in South Africa very soon after the deregulation of the insurance market in 1994. While the market share of HMO-like contracts is about 4\%, MSA-based contracts are rather widespread with 51\%.

Yip and Hsiao (1997) describe the experiences of the Chinese with a MSA-based social insurance experiment in two cities, and conclude that there is only little empirical evidence of the impact of MSAs on the total health care expenditures.

In the USA advantages and disadvantages of MSAs are discussed at the political and academic level.\footnote{One can mention: American Academy of Actuaries (1995a, 1995b), Bunce (2001), Friedman (2001), Moon et al. (1996), Ozanne (1996), Pauly and Herring (2000), Pauly and Goodman (1995), Scandlen (1998), Stano (1981), and Zabinski et al. (1999), but the list does not claim to be complete.} Schreyoegg (2003) discusses the eventualities of a MSA introduction in Germany. However, there is only little theoretical insight in the decision-making process of an individual when setting up a MSA. Heffley and Miceli (1998) analyze MSAs only as an example within their framework of incentive-based health care plans.

This paper shall shed some light on the pros and cons of a medical
savings account compared to a deductible insurance with no dynamic bonus options. In the first part, we will use a two-period expected utility model to answer the question whether a risk averse individual would prefer a MSA over a common deductible insurance, and under which circumstances.

In the second part, we will use Swiss health insurance data to simulate several versions of MSAs and a deductible insurance in order to test the predictions of the theoretical model. Furthermore, we search for the insurance plan which maximizes the insured’s individual expected utility and the average welfare of the entire sample.

This simulation extends thoroughly the contribution of Eichner et al. (1997) who have done a similar estimation with American firm level data. While they have focused on the distributional aspects of an introduction of medical savings accounts, we consider explicitly the welfare effects and compare rather different versions of MSAs like the MSA proposed by Stano (1981) and a rebate policy.

2 The Theoretical Model

Two types of individual health accounts based insurances can be distinguished: the US-type and the Singapore-type MSA. The main difference lies in the flexibility of the overall deductible: while this one is fixed for US-type MSAs, the maximum of one’s own contribution to the health care costs under a Singapore-type MSA depends on the balance of the MSA. The above mentioned Archer MSA and the MSAs in South Africa belong to the US-type MSAs. Medisave in conjunction with Medishield and Medifund in Singapore can be interpreted as a MSA with a flexible overall deductible: the higher the savings in the Medisave account the higher can be the insured’s copayments. Only, if the Medisave account is depleted - because of the 20% copayments of the Medishield insurance - and the insured is not able to cover the remaining costs out-of-pocket, Medifund will step into the breach.⁵ The Chinese MSA and the individual health accounts proposed by Stano (1981) belong also to the Singapore-type MSA. Both approaches consist of individual MSAs and a social insurance which covers the health care costs in the case of a depleted MSA.

Differences in the flexibility of the overall deductible lead directly to differences in the regularity of the insured’s contribution into her MSA. If the overall deductible is fixed, there is no need to require regular deposits. It can be left to the insured herself to decide on. However, if the overall deductible depends on the balance of the account, the premium of the additional insurance will depend on it, too. Therefore, in order to calculate an expected

---

⁵This is, admittedly, a highly compressed representation of Singapore’s health care financing system. For a detailed description, we want to refer to Massaro and Wong (1996) and the Internet site of Singapore’s Ministry of Health: http://www.gov.sg.
balance of the MSA, one need regular deposits into the account. This can be observed in Singapore and China where the insureds must pay a certain share of their monthly income into the MSA.

What we want to model here is the Singapore-type MSA. The reason is that a US-type MSA is not very much different from a standard deductible insurance due to the fixed overall deductible.

For simplicity, we use a two-period model. These periods can be either interpreted as policy-years or as phases of one’s life like youth and age. At first, we define the risk to be insured against.

**Assumption A. 1 (The Risk).** Let \( \Omega \) be the set of all possible diseases.

1. \( A : \Omega \to \mathbb{R}^+ \) is a continuous random variable,
2. \( F(a_1, a_2) \), with \( a_1, a_2 \in A \), is a joint distribution function where the index indicates the period.
3. \( F \) is common knowledge and insurable.

The elements of \( A \) can be interpreted as the whole follow-up costs of a disease. Thus, \( F \) is the joint distribution function of the insured’s medical expenses in period one and two. With insurable, we mean that, firstly, there is a real risk involved, i.e., \( F \) is not degenerated, and, secondly, an insurance policy against the losses born by \( F \) would not be too expensive, and a potential insurance buyer can afford the policy.

**Assumption A. 2 (Individual’s Preferences).** Let \( u \) be the insured’s utility function, and \( w \in \mathbb{R}^+ \) the insured’s wealth with \( u : w \to \mathbb{R} \). Furthermore, denote the discount factor with \( \rho \in (0, 1] \).

1. \( u \in \Upsilon := \{ u : u' > 0, u'' < 0 \ \forall w \in \mathbb{R}^+ \} \),
2. \( v := \mathbb{E}(u_1 + \rho u_2) \), where the index only indicates the period, without altering the utility function itself.

Firstly, we assume a strict risk averse individual whose welfare per period can be represented by a continuous, strict concave, and two times differentiable utility function. In addition, the insured’s utility is supposed to depend on her wealth, only. So, for simplicity, we ignore the possible impact of an illness on the insured’s welfare. This, however, should favor the deductible insurance because of the simultaneous occurrence of out-of-pocket payments and illness. Secondly, we assume that the preferences over different states of the world, i.e., as a consequence of different insurance policy parameters, can be represented by the sum of expected utilities per period.

---

6Similar single-period models of a deductible insurance can be found in Eeckhoudt et al. (1991), Gould (1969), Mossin (1968), and Schlesinger (1981).

7See Zweifel (1992) for a similar argumentation.
The supply side is, for the sake of simplicity, only rudimentary modeled: An insurance company offers the insurance buyer a policy which yields zero expected profits; either due to competition or because of public regulations. Thus, the insurance buyer can be insured according to a premium function which maps the expected indemnity payments and costs of the insurance company to the policy premium.

**Assumption A. 3 (Premium Function).** Let \( p \) be the premium per period and \( i \) the indemnity payments born by the policy. In addition, denote the insurance loading with \( \lambda \geq 0 \) and the interest factor with \( r \geq 1 \). Then, the premium function is

\[
p = \frac{1 + \lambda}{1 + r} E(i_1 + i_2),
\]

with \( p \) constant in period one and two.

In a single-period model, the premium function would be \( p = (1 + \lambda)Ei \). The denominator in (1), therefore, equalizes the per period premiums. In addition, one has to include the interest costs or gains, respectively, of the indemnity payments in period one. Hence, these have to be multiplied by the interest factor.

The assumed premium function has the advantage that it already incorporates increasing medical costs in older years. Due to the fact that the premium shall be of equal size in each period, one would have to pay a higher premium in younger years - compared to the actual incurred costs - than in older years. Thus, the insurer can accumulate enough resources in the first years in order to be able to pay the higher costs when the insured gets older. However, this means that the incentive to save on a MSA is not related to the desire to avoid age-related increases in premium payments because this is already assured by the premium function.

**The Policies**

With a common deductible insurance, we mean an insurance where, in each period, the insured pays the medical expenses up to the amount of the deductible. All costs above are covered by the insurer. The deductible is fixed and from equal size in each period.

A MSA is a combination of a savings account and an insurance. The insured agrees to pay a certain amount of money into the MSA in each period. Withdrawals from the MSA are only allowed for medical purposes. Medical expenses are firstly paid out of the MSA. If the savings are not enough, and a further out-of-pocket maximum is reached, the insurer pays the rest. After the policy runs out, the remaining savings are distributed to the insured.

It is rather cumbersome to compare two different policies directly. That’s why, we define a combined policy with two parameters: the maximum out-of-
pocket payment per period (MOP) \( d \geq 0 \) and the rate of saving \( \delta \in [0, 1] \) of unused MOP which will be saved in the MSA. So, a chosen \( \delta = 0 \) means the policy is a common deductible insurance. And \( \delta = 1 \) means the insurance buyer chooses to pay all copayments out of the MSA. However, a rate of saving between zero and one is also possible and would include an additional out-of-pocket payment after the savings account is depleted. In each policy year the timetable shall be as follows:

1. The medical expenses of the period become known and are withdrawn from the MSA.
2. If the savings are not enough, the insured has to pay out-of-pocket.
3. If the costs exceed the MOP, too, the insurer pays the remaining costs.
4. If the medical expenses are lower than the MOP, the insured pays the share of unused MOP into the MSA according to the rate of saving.
5. In the last policy year, the remaining savings will be paid back to the insured.

Table 1 shows how such a combined policy would work under different rates of saving. In the example it is assumed that in the first year medical costs of 200 and in the second of 1500€ occur. With \( \delta = 0 \), no savings take place. Therefore, the MSA balance is zero. However, if \( \delta = 0.5 \), the insured saves the half of the unused MOP \((0.5 \times (1000 - 200))\) which, then, can be used to partially pay the costs of the second period. As one can see, with an increase of the savings rate, the indemnity payments paid by the insurer shrinks. Therefore, the lowest premium would be paid by an insured with a \( \delta = 1 \). In our example, this premium would be zero.

<table>
<thead>
<tr>
<th></th>
<th>Rate of Saving</th>
<th>( \delta = 0 )</th>
<th>( \delta = 0.5 )</th>
<th>( \delta = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>Policy Year</td>
<td>1. 2.</td>
<td>1. 2.</td>
<td>1. 2.</td>
</tr>
<tr>
<td>3.</td>
<td>MSA Balance, 1. Jan</td>
<td>0 0 400 0</td>
<td>0 800</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Medical Expenses</td>
<td>200 1500 200 1500</td>
<td>200 1500</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Indemnity payments</td>
<td>0 500 0 100 0</td>
<td>0 0</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>MSA Balance, 31. Dec</td>
<td>0 0 400 0 800 300</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( a = \max\{4. - 3. - d, 0\} \).
\( b = \max\{\delta(3. + d - 4.), 0\} \). If \( 4. \geq 3. \), \( 6. = \max\{\delta(3. + d - 4.), 0\} \). If \( 4. < 3. \), \( 6. = 3. - 4. + \delta d \).
To abstract from other savings strategies, we impose two additional assumptions:

**Assumption A. 4.** Let \( y \) be the insured’s income per period. The insured’s income is constant over time, i.e., \( y_1 = y_2 = y \).

**Assumption A. 5.** \( \rho r = 1 \).

Denote with \( l \) (low) the case if the insurer’s indemnity payments in the period are zero due to relative low medical expenses, and with \( h \) the opposite. Therefore, in a two-period model, we can distinguish four different disease histories: \( ll, lh, hl \) and \( hh \).

**Figure 1:** Possible disease histories in a two-period model.

In Figure 1, each point reflects a pair of medical costs in period one and two. In the case of \( ll \), the insured has to bear all of the expenses. This area increases with \( \delta \).

Now, we can calculate the expected utility in the case of a certain disease history:

\[
\begin{align*}
    v_{ll} &= \mathbb{E}_l[u(y - p - a_1 - \delta(d - a_1)) + \rho u(y - p - a_2 + r\delta(d - a_1))], \\
    v_{lh} &= \mathbb{E}_{lh}[u(y - p - a_1 - \delta(d - a_1)) + \rho u(y - p - a_2)], \\
    v_{hl} &= \mathbb{E}_{hl}[u(y - p - a_1 - \delta(d - a_1)) + \rho u(y - p - d)], \\
    v_{hh} &= \mathbb{E}_{hh}[u(y - p - d) + \rho u(y - p - d)].
\end{align*}
\]  

(2)

The insured’s overall expected utility is, therefore, the sum of the utility of all possible states of the world

\[
v(d, \delta) = v_{ll} + v_{lh} + v_{hl} + v_{hh}.
\]  

(3)
At last, we assume that the insurance loading – and, therefore, the costs of the insurer – is positive and independent of the rate of saving.

**Assumption A. 6.** \( \lambda > 0 \) constant \( \forall \delta \).

Thus, the premium shall not be actuarial fair. This gives an incentive to take no full coverage. Furthermore, the discrimination between MSA and DI is not supposed to be based on different cost structures.

**The Expected Utility Criterion**

In this section, we want to compare MSA and DI from the insured’s point of view. Can there be any reason found why a potential insurance buyer will choose a MSA over a DI? The problem for an expected-utility maximizer when facing the choice of a combined policy is as follows:

\[
\begin{aligned}
\max_{d, \delta} v(d, \delta) &= \mathbb{E}(u_1 + \rho u_2) \\
\text{subject to } p(d, \delta) &= \frac{1 + \lambda}{1 + r} \mathbb{E}(r_1 + i_2). 
\end{aligned}
\]

(4)

Now, we can answer the question whether a MSA is superior to a DI or vice versa: if the insurance buyer chooses \( \delta^* = 0 \), she prefers the pure deductible insurance, where the * denotes the solution to problem (4). If she demands a \( \delta^* > 0 \), she prefers a MSA. However, you have to know the utility function and the distribution of medical costs to predict the optimal \( \delta \) directly. Instead of specifying this here, we will follow the question for what types of insurance buyers one can expect a choice of \( \delta^* = 0 \) and for whom not. Later, in the simulation, we will specify the parameters and can then directly compare both policies.

Let us firstly show that an insurance buyer will always choose a positive maximum out-of-pocket payment per period whenever the insurance premium is not actuarial fair.

**Lemma 1 (Positive MOP).** Suppose A.1-A.6; then \( d^* > 0 \).

**Proof.** The lemma is proved if the first derivative of \( v(d, \delta) \) w.r.t. \( d \) at the point \( d = 0 \) is positive for every \( \lambda > 0 \). Note that with a zero MOP only \( hh \) is possible and \( u'_1 = u'_2 = u'(y - p) \) in all situations.

\[
v'_d|_{d=0} = -p'_d \mathbb{E}(u'_1 + \rho u'_2) - \mathbb{E}_{hh}(u'_1 + \rho u'_2),
\]

(5)

with

\[
-p'_d = \frac{1 + \lambda}{1 + r} \mathbb{E}_{hh}(1 + r).
\]

(6)

Let \( v' := \mathbb{E}(u'_1 + \rho u'_2) \). Setting (6) into (5) yields \( v'_d|_{d=0} = \lambda v' > 0 \) if and only if \( \lambda > 0 \). Therefore, the expected utility at \( d = 0 \) cannot be maximal. \( \square \)
Now, let \( D_0 := \{ d : v'_{d} \geq 0 \text{ and } \delta = 0 \} \) and \( D_0^* := \{ d : \text{maximal and } \delta = 0 \} \). Then, clearly \( D_0^* \subset D_0 \). Below, we will need the following result:

**Lemma 2.** Suppose A.1-A.6 and let \( u'_d = u'(y - p - d) \); then, for every \( d \in D_0 \): \( v \frac{1 + \lambda}{1 + r} u' \geq v'_d \).

**Proof.** Setting \( v'_d|\delta=0 \geq 0 \) and multiplying by \( r \) yields the result. See appendix A.1 for the details. \( \square \)

With that in mind, we can formulate a condition such that for every positive \( d \in D_0 \) the insured can increase her welfare when choosing a positive rate of saving.

**Lemma 3 (Positive Rate of Saving).** Suppose A.1-A.6; then, for every given positive deductible \( d \in D_0 \), a sufficient condition for \( \delta^* > 0 \) is

\[
\mathbb{E}_d[(d - a_1)(u'_1 - r\rho u'_2)] \leq \mathbb{E}_{lh}[(d - a_1)(u'_d - u'(y - p - a_1))] \tag{7}
\]

at the point \( \delta = 0 \).

**Proof.** In the same manner as before, we are done if we can show that \( v'_\delta > 0 \) at the point \( \delta = 0 \) when (7) applies.

\[
v'_\delta|\delta=0 = -p'_\delta v' - \mathbb{E}_d[(d - a_1)(u'_1 - r\rho u'_2)] - \mathbb{E}_{lh}[(d - a_1)u'_1], \tag{8}
\]

with

\[
-p'_\delta = \frac{1 + \lambda}{1 + r} \mathbb{E}_{lh}[r(d - a_1)]. \tag{9}
\]

Setting (9) into (8) yields

\[
v'_\delta|\delta=0 = -\mathbb{E}_d[(d - a_1)(u'_1 - r\rho u'_2)]
+ \mathbb{E}_{lh} \left\{ (d - a_1) \left[ \frac{1 + \lambda}{1 + r} v' - u'_1 \right] \right\}. \tag{10}
\]

Because of Lemma 2 and \( r\rho = 1 \), equation (10) can be transformed to

\[
v'_\delta|\delta=0 \geq -\mathbb{E}_d[(d - a_1)(u'_1 - u'_2)]
+ \mathbb{E}_{lh} \left\{ (d - a_1) \left[ u'_d - u'(y - p - a_1) \right] \right\}. \tag{11}
\]

Due to risk aversion \( v'_\delta > 0 \) at the point \( \delta = 0 \) if (7) applies, and the insured can increase her expected utility when choosing a positive rate of saving. \( \square \)

This result leads us directly to our first main statement:

**Proposition 1.** Suppose A.1-A.6; then, if (7) applies, a risk averse individual will always prefer a MSA over a pure deductible insurance, i.e., \( d^* > 0 \) and \( \delta^* > 0 \).
Proof. This result follows immediately from the three lemmas above: Due to Lemma 1 \( d^* > 0 \); together with (7) and the fact that \( d^* \in D_0^* \subset D_0 \) all of the conditions for \( \delta^* > 0 \) are met.

The reason for this result is the relative high income in period one if one chooses a pure deductible solution and \( lh \) occurs. With a positive rate of saving, the insured can transfer some income from \( lh \) to other states of nature, i.e., where she has to bear more medical costs. There is no way to do that with a change of the deductible itself. An increase of \( \delta \), however, decreases the running income for consumption in the first period, immediately. This leads to a lower premium, which increases the insured’s utility in all situations.

In the case of \( ll \), a rate of saving higher than zero leads only to wealth shifting from the first to the second period. This fact is represented by the lhs of (7). If the medical costs are independent and identical distributed, there is no difference between the marginal utilities in period one and two. That’s why, the second term in (7) becomes negative, how the following corollary states:

**Corollary 1.1.** Suppose A.1-A.6 and independent and identical distributed medical costs in period one and two; then, a risk averse individual will ever prefer a MSA over a pure DI, formally

\[
A.1 \text{ – } A.6, \quad f(a_1, a_2) \equiv f(a_1)f(a_2) \quad \& \quad f(a_1) = f(a_2) \quad \Rightarrow \quad \delta^* > 0 \quad (12)
\]

Proof. See appendix A.2.

Only, if there is a higher weight on those cases where the medical costs in the first period are higher than in the second period equation (7) could not be met. In this case, the marginal costs of an increase of \( \delta \) in the first period could be higher than the marginal gain in period two. If this effect is high enough, the above mentioned utility gain through consumption smoothing could be overwhelmed.

Until now, we have only seen that the rate of saving can be positive under certain circumstances. But, we do not know if a rate of saving of 50% or even 100% is optimal. One way to show that the latter will be preferred to all other rates of saving is to ensure that the expected utility is ever increasing in \( \delta \).

**Proposition 2.** Suppose A.1-A.6; then, if the probability of \( lh \) is sufficiently higher than the probability of \( ll \), a risk averse individual will ever prefer a pure MSA with \( \delta^* = 1 \).

Proof. If we show that \( v'_\delta > 0 \) for all \( \delta \in [0, 1] \) and possible candidates for \( d^* \), we are done because then \( \delta^* \) will be increased until its upper boundary. This can be done by using a little altered version of Lemma 2 and 3. The details can be found in the appendix A.3.

\[ \Box \]
The relation between the probabilities of the illness histories \( ll \) and \( lh \) is crucial for the decision between a classic deductible insurance and a MSA. If \( lh \), a MSA smooths the consumption stream better. However, in the case of \( ll \), a MSA leads to relative low utility in the first period and a higher one in the second period due to the saving of the deductible. In contrast, a classic deductible insurance requires only to pay the medical costs in both periods, and the utilities do not differ as much. When \( hl \) or \( hh \), the two policies do not differ in their associated consumption streams, thus, the probability of \( hl \) and \( hh \) should not matter.

A MSA allows to accumulate capital in healthy years which can be used later on – in years of sickness – to keep the premium affordable. If we interpret the periods as the youth and the age, the model predicts a positive rate of saving due to the fact that the medical costs raise in the second part of one’s lifetime for most of us, and, therefore, \( lh \) is more likely.

Arrow’s Theorem

The use of Arrow’s (1963) theorem allows another way to think about the superiority of a policy. The theorem states that a policy cannot be optimal if the insured gets indemnity payments in a state of nature where the insured’s wealth in another state of nature is lower. With a DI, this would be exactly the case when \( lh \) and \( hl \) because the insured’s wealth is in \( hh \) lower than in all the other states. Thus, in a two-period model a DI cannot be optimal. On the other hand, with a pure MSA, the insured’s wealth in \( lh \) and \( hh \) is equal, and, therefore, the welfare should be higher. However, a pure MSA is also not optimal because the wealth in \( hl \) is higher than in \( lh \) and \( hh \). The only optimal two-period policy would be a fixed overall deductible over the sum of the medical expenses of the two periods. This can be done, for instance, if the insured can go in debt on her MSA. However, how Cohen (2001) shows, such an aggregate deductible can lead to an increased moral hazard problem if the deductible is reached in the first period, and might, therefore, not be preferable.

3 Simulating Medical Savings Accounts and Deductible Insurance

The idea of the simulation is to extract from real life data a reasonable approximation of a possible medical expenditure history. Two main attributes have to be considered. First, the distribution of medical expenditures among the insureds is enormous unequal. As we will see in our sample we find that 80% of the costs in one policy year are born by roughly 20% of the insureds. And the second is that the expenditures in one year are highly correlated with the expenditures in the previous years. We will take both
into account. The procedure we will follow is an extension of that used by Eichner et al. (1997). At first, we estimate with the help of a three-stage regression model several parameters which we then use to perform the simulation. Eichner et al. (1997) don’t take into account the possibility that an insured dies. This implies, however, that they cannot simulate beyond a certain age. We instead estimate an additional dying probability which allows us to extrapolate far more longer.\(^8\)

3.1 The Data

The simulation is based on a sample from the Swiss health insurance company CSS. This is a rather comprehensive dataset and includes individual health expenditures over three years (1997 through 1999). The insureds are from the cantons Zurich and Geneva. We only include those subjects who were insured throughout the whole 24 months in 1997 and 1998. In order to estimate a probability of death, we allow an ending of the policy during the year of 1999 where we could observe whether the reason was the insured’s death or not. This reduces our sample to 121,428 individuals. 55.67% of the sample are female and nearly 1% died in 1999. Table 2 shows the descriptive statistics for the age and the medical expenses.

<table>
<thead>
<tr>
<th>Table 2: Descriptive Statistics of the Swiss Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=121428</td>
</tr>
<tr>
<td>Age in 1999</td>
</tr>
<tr>
<td>Expenses in 1997 (€)</td>
</tr>
<tr>
<td>Expenses in 1998 (€)</td>
</tr>
<tr>
<td>Expenses in 1999 (€)</td>
</tr>
</tbody>
</table>

Since 1996, basic health insurance in Switzerland has been mandatory for all citizens. The premium is for all insureds equal and depends solely on the chosen deductible. The deductibles levels are 230 CHF\(^9\), 400, 600, 1200 and 1500. Beyond the deductible, a coinsurance of 10% is required up to a maximum-out-of-pocket limit which is the deductible plus 600 CHF. The sample includes the information about the chosen deductible. The lowest deductible was chosen by 66.2%. 21.4% chose the 400 CHF deductible, and 5.7, 2 as well as 4.4% went with the highest copayments.

\(^8\)We will simulate until the age of 100 where we then assumes everybody will die. The oldest insured in Eichner et al. (1997)’s model was 65 which is certainly due to the fact that they explicitly simulated with the situation of an American company in mind. After the age of 65, the elderly get social health insurance by Medicare.

\(^9\)Until 1998 this amount had been 150, where 1 CHF \(\approx 0.63\) €
Specialities of Medical Expenditures

As we mentioned above, persistence and high correlations between medical expenses in consecutive years have to be taken into account if one wants to simulate medical histories.

Figure 2: Lorenz curve for the Distribution of Medical Expenditures in 1999

Figure 2 shows the Lorenz curve of the distribution of medical costs in the year 1999. As one can observe 76% of the cumulated medical costs are incurred by only 20% of the insureds. In addition, roughly 19% of the subjects filed no medical claim in 1999 at all, and 50% had less than 500 €. However, one has to take into account that expenses below the margin set by the deductible are not observable because the insureds have no incentive to report them. This is a problem which we discuss below when we introduce the model. What is important here for our simulation is that we have to model the decision to visit a physician explicitly in order to explain a large share of non-positive medical expenses. In addition, the strong unequal distribution of the costs among the insureds requires to take the logarithmic value in order to get an approximate normal distribution.

From the age of 80 on, the picture changes. In 1999, 59% of the cumulated medical expenses of the insureds older than 80 years are born by 20%. For the 90 years old and above this number decreases further to 45%. Therefore, with a certain age, the distribution of the medical costs gets more and more equal. Health related problems lose their accidental character and get common among the insureds. We take this into account and divide the
estimation of the expected medical costs into two different regression functions: one for the insureds younger than 80 years and one for the older. For the latter, we use a non-logarithmic form.

One will notice the strong correlation of individual medical expenditures below when we present the results of the estimation procedure. However, as Eichner et al. (1997) also have done, we can get an impression of the implied persistence in the data if we contemplate the average medical costs in consecutive years sorted by the observed deciles of the distribution in 1997 (Figure 3).

![Figure 3: Mean Medical Expenses by 1997 Deciles](image)

Similar to the findings of Eichner et al. (1997), we can conclude that individuals with high medical costs in 1997 are very likely to have high expenses in the following years as well. Thus, we will treat the expenses of 1997 and 1998 as explanatory variables for the estimation of the costs in 1999.

A third fact is that the average medical expenses increase over the lifetime. This is captured in figure 4. Noticeable is that the relationship is far from linear, thereby we will include a squared age variable in our regression analysis. Eichner et al. (1997) only consider individuals not older than 65 in their simulation of medical savings accounts. The figure, however, suggests that shortly after the age of 65 is reached a tremendous rise in medical costs occur. Therefore, simulating beyond this age will yield a better basis for an examination of MSAs.
3.2 The Model

The in the previous section discussed specialties of individual medical expenditures data lay the ground for our regression model. To estimate the expected medical costs, we use a two-stage model. At the first stage, the individual decides whether to visit a physician or not. These considerations shall be based on the demographic characteristics gender and age.\(^{10}\) During the validation of the model, we’ve concluded that age has to be included up to the third power. In addition, the chosen copayments level\(^{11}\) and whether the insured has gotten a premium reduction\(^{12}\) are supposed to have an impact.

At last, we assume that the decision to visit a physician is determined by the medical history of the insured. This is captured by the medical costs of the previous years \((m7 \text{ and } m8)\) and the three dummy variables which are 1 if the individual had only positive medical expenses in 1997 \((v7)\), only in 1998 \((v8)\), or in both years \((v78)\). We use a probit regression model because we cannot directly observe the latent variable behind this decision.

Thus, individual \(i\)’s probability of visiting a physician \((\pi_i)\) or causing

\(^{10}\) sex = 0 female, \(sex = 1\) male, age in the year 1999 measured in years.

\(^{11}\) \(ded \in \{0, 1, 2, 3, 4\}\) where 0 is a deductible of 230 and so on.

\(^{12}\) If the income of the insured is below a certain threshold, a premium reduction can be granted. This variable can, therefore, be interpreted as a measure of the insured’s income. See also Werblow and Felder (2002). \(p = 1\) premium reduction, \(p = 0\) no reduction.
positive medical costs, respectively, shall be estimated by

\[ \pi_i = F(x_i, b) = F(\alpha_1 + (sex_i + age_i + age_i^2 + age_i^3)'\beta_1 + (ded_i + p_i)'\gamma_1 + (m7_i + m8_i + v7_i + v8_i + v78_i)'\delta_1) \]

where \( F \) is the standard normal distribution function.

At the second stage, the amount of incurred medical costs \((m9)\) is determined for those who decided to attend a physician. However, we have to take into account the selection bias because only those with positive medical costs are considered. This can be done by including an additional independent variable \( \lambda_i \) which is calculated from the fitted values of the probit regression. In addition to the variables used in the probit regression, we include the squared values of the medical costs of the previous years as well. Furthermore, we add a dummy variable \( age50 \) which is 1 for all insureds not younger than 50. As mentioned above, we estimate two equations: one for those younger than 80 years and one for the older individuals. For the former, we use a log-linear OLS specification, for the latter a linear OLS. Thus, we get

\[ \ln m9_{i|age<80 \& m9>0} = \alpha_2 + (sex_i + age_i + age_i^2 + age_i^3 + age50)'\beta_2 + (ded_i + p_i)'\gamma_2 + (m7_i + m8_i + m7_i^2 + m8_i^2) + v7_i + v8_i + v78_i)'\delta_2 + \hat{\lambda}_i\rho_2 + \epsilon_2, \]

and

\[ m9_{i|age\geq80 \& m9>0} = \alpha_3 + (sex_i + age_i + age_i^2 + age_i^3)'\beta_3 + (ded_i + p_i)'\gamma_3 + (m7_i + m8_i + m7_i^2 + m8_i^2) + v7_i + v8_i + v78_i)'\delta_3 + \hat{\lambda}_i\rho_3 + \epsilon_3. \]

Zero costs in the sample does not necessarily mean that the insured caused no medical expenses in this year. Because of the deductible, the individuals will not report a medical treatment until she can expect to get reimbursed by the insurer. In addition, the different deductible levels might generally lead to a self selection: insureds who expect to have relative low medical costs in the next year could choose a high deductible. Therefore, the estimated parameters of \( ded \) cannot be simply interpreted as an evidence for moral hazard. Our approach is different because we are not interested in

\footnote{\( \lambda_i = f(x_i, b)/F(x_i, b) \) where \( f \) is the density function of the standard normal distribution. For a detailed description see Werblow and Felder (2002) or Maddala (1983).}

\footnote{For a study which explicitly recognizes the self selection see Werblow and Felder (2002). Furthermore, Manning et al. (1987) estimate the impact of copayments on the demand for medical care with data of the RAND Health Insurance Experiment where self selection has been ruled out.}
measuring the effect of moral hazard. We estimate the parameter of \( \text{ded} \) in order to catch possible differences in risk and/or attitudes toward medicine. Only after the simulation, we will adjust the obtained medical expenditures for moral hazard. The procedure will be described below.

In the simulation of medical expenditures for individuals until the age of 100, we have to include the possibility of an insured’s dead. Instead of taking the values out of a mortality table, it is desirable to estimate these on the basis of our sample. This enables us to specify certain parameters of the impact of demographic, income, and medical history related variables. We use again the probit model because we can only observe the event of dead. The following equation is estimated:

\[
\tau_i = F(x_i + c) = F(\alpha + (sex_i + age_i + age_i^2)\beta_1 + (\text{ded}_i + p_i)\gamma_4 + (m8_i + v8_i)\delta_4).
\]

The hypothesis that the mortality is only influenced by the medical history of the preceding year could not be ruled out. As a consequence, we only include \( m8 \) and \( v8 \).

<table>
<thead>
<tr>
<th>( \pi_i )</th>
<th>( \ln m9_i(&lt;80) )</th>
<th>( m9_i(\geq80) )</th>
<th>( \tau_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>1.420 e - 01 ***</td>
<td>1.668 e - 02 ***</td>
<td>8.938 e + 05 **</td>
</tr>
<tr>
<td>( sex )</td>
<td>-2.094 e - 01 ***</td>
<td>2.224 e - 01 ***</td>
<td>3.222 e + 02</td>
</tr>
<tr>
<td>( age )</td>
<td>-2.365 e - 02 ***</td>
<td>6.774 e - 02 ***</td>
<td>-3.094 e + 04 **</td>
</tr>
<tr>
<td>( age^2 )</td>
<td>4.704 e - 04 ***</td>
<td>1.142 e - 03 ***</td>
<td>3.546 e + 02 **</td>
</tr>
<tr>
<td>( age^3 )</td>
<td>-2.154 e - 06 ***</td>
<td>6.564 e - 06 ***</td>
<td>-1.348 e + 00 **</td>
</tr>
<tr>
<td>( age50 )</td>
<td>3.747 e - 02 *</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{ded} )</td>
<td>-1.049 e - 01 ***</td>
<td>1.236 e - 01 ***</td>
<td>-2.127 e + 02</td>
</tr>
<tr>
<td>( p )</td>
<td>1.855 e - 02</td>
<td>4.803 e - 02 ***</td>
<td>5.979 e + 02 ***</td>
</tr>
<tr>
<td>( v7 )</td>
<td>5.049 e - 01 ***</td>
<td>-9.883 e - 01 ***</td>
<td>-2.982 e + 03 **</td>
</tr>
<tr>
<td>( v8 )</td>
<td>6.233 e - 01 ***</td>
<td>-1.412 e + 00 ***</td>
<td>-3.335 e + 03 *</td>
</tr>
<tr>
<td>( v78 )</td>
<td>1.303 e + 00 ***</td>
<td>-2.557 e + 00 ***</td>
<td>-4.515 e + 03 **</td>
</tr>
<tr>
<td>( m7 )</td>
<td>1.082 e - 04 ***</td>
<td>4.570 e - 05</td>
<td>2.225 e - 01 ***</td>
</tr>
<tr>
<td>( m8 )</td>
<td>2.357 e - 04 ***</td>
<td>8.941 e - 05 ***</td>
<td>6.200 e - 01 ***</td>
</tr>
<tr>
<td>( m7^2 )</td>
<td>-2.161 e - 10</td>
<td>-1.879 e - 06</td>
<td></td>
</tr>
<tr>
<td>( m8^2 )</td>
<td>-8.991 e - 10 ***</td>
<td>-5.583 e - 06</td>
<td></td>
</tr>
<tr>
<td>( \lambda )</td>
<td>2.436 e + 00 ***</td>
<td>2.407 e + 03</td>
<td></td>
</tr>
</tbody>
</table>

The obtained coefficients of the estimation procedure are summarized in Table 3. These results seem reasonable. Men are likely to visit the physician more seldom than women. But, they cause more medical costs and die earlier. The impact of age is quite complex in this model due to

\[
R^2 = 0.3679 \quad \text{0.4192} \quad \text{0.3929} \quad \text{0.3235}
\]
the amount of age related variables. In general, physician visits are likely in the early and later years of a lifetime. The costs increase with the age, even with exponential rate in older years. The large estimated intercept of equation (15) compared to (14) is good sign for the structural break around this age.

The influence of the level of deductible (ded) is negative concerning physician visits but positive on the level of expenditures. One explanation is indeed that a high deductible might lead to a postponement of the decision to seek medical help. In consequence, a clustering of physician visits in a given policy year could than increase the costs of a treatment. However, these signs are also explained by the fact that if the insureds cause costs lower than their amount of deductible, they simply do not report them. Therefore, even if two individuals have equal expenses, the one with the higher deductible reports less often but higher amounts. Interestingly, the deductible shows a significant negative impact on the probability of death. This seems to favor the argument that risk selection takes place.

A physician visit in a preceding year (v7, v8, v78) has a positive effect on the probability to seek medical treatment in the next year. However, as the model suggests, this can generally decrease the amount of medical expenses. Therefore, a continuous treatment might reduce the costs which on the other hand could explain the negative impact of the deductible level on the medical expenditures. But, the impact of the size of the medical costs is unambiguous: the larger the costs in the preceding years the higher the probability of a physician visit and the larger the expected medical costs in the next period. This is exactly the above described persistence which is clearly supported by our model.

The Fit of the Simulation Procedure

In principle, we follow the procedure of Eichner et al. (1997), however, we add the possibility that the insureds can die. The estimated parameters allow us to predict for each year and individual the probability of visiting a physician, the size of the expected medical expenditures if positive, and the probability of death. But, in order to capture the actual distribution of the data as close as possible, we add to each predicted value a (pseudo) residual drawn out of a sample which is grouped by gender, age, and medical expenses of the preceding years. In particular: on the first stage, we predict π, for each individual and add a (pseudo) residual.15 If this value is greater than 0.5, we assume the individual seeks medical treatment, and predict the amount of the incurred medical costs according to equation (14) or (15), respectively, and add again a residual. Otherwise the medical costs are set to zero in this year. Furthermore, we calculate τi - the probability of death.

15Because π, is estimated by a probit regression, we don’t know the ”real” residual which would be the difference between predicted and actual value of the latent variable.
After adding a pseudo residual, we assume that the individual dies at the end of the year if the obtained $\tau_i$ is greater than 0.5. The predicted expenditures of the current year become then the independent variable of the following year, and the procedure gets reapplied.

**Figure 5:** Actual vs. Predicted Medical Expenses by Age

![Actual vs. Predicted Medical Expenses by Age](image)

To establish whether this procedure delivers a reasonable approximation of the real data, we will compare the obtained values with their actual counterparts. In order to do so, we have started the simulation with all the individuals of the sample who were 25 years old in the year of 1999. For those 1427 insureds, we have repeatedly applied the above described predictions for 76 years until they reached the age of 100.

First of all, we compare average annual medical expenses ordered by age. This shows us whether the magnitude of the predicted expenditures is acceptable. Figure 5 does the comparison, where the dashed line always represents the predicted values. Up to the age of 95, the two lines lie very close to each other. Especially, the switch from the log-linear to the linear regression at the age of 80 has assured this excellent fit. However, after the age of 95, we noticeably underestimate the real data. This is due to the fact that i) the sample size of the actual data becomes very small near the age of 100, and ii) we have underestimated the mortality rate. The latter can be verified by Figure 6. All dots above the zero line stand for a predicted mortality rate lower than the actual. The fit is quite good until the age of 85. After that, the simulation procedures delivers a mortality rate which is up to 10% smaller than the real one. Therefore, in our simulation 54 individuals
reach the age of 100, where only 10 insureds in the actual data are of that age. This leads to a more extreme distribution of medical expenditures among the actual elderly in the sample which cannot be captured by the simulation.

Another comparison shall clarify whether the simulation has caught the distribution of medical expenditures among the insureds in a reasonable manner. Figure 7 demonstrates this for two age groups: the 30 and the 60 years old. Again, the dashed lines stand for the simulated values. The figure indicates that from expenses of 1000€ on, the fit is acceptable. However, below this margin the simulation shows in the case of the 30 years old a overestimation of the probability of physician visits and a underestimation for the 60 years old. This is a systematic pattern in the way that the overestimation of $\pi_i$ gets reduced with increasing age. Between the age of 35 till 55 the lines are almost identical. In other words, the distribution of medical expenses among the insured of the simulation is slightly more extreme in younger years and less skewed in older years than the actual distribution. Although this pattern is systematic, we could find no way to improve the procedure. However, we believe that the magnitude of this problem is low enough to continue the analysis.
3.3 Results of the Comparison

The Simulated Policies

We are going to compare three different MSA plans and a deductible insurance. Table 4 summarizes the policies. First of all, a MSA.ded is identical to the medical savings account analyzed in the theoretical part of this paper. It features a certain deductible and a specified savings rate. Each year, the insured pays medical expenses out of the interest bearing MSA. If the balance is not enough to cover the costs, the policy holder has to pay out-of-pocket up to the amount of the deductible. Beyond that the insurer pays the remaining costs. Only if the deductible is not fully used up in a year, a share of the remaining amount is contributed into the MSA according to the savings rate. However, in contradiction to the theoretical model, there is no certain end of the insurance after that the policy holder would withdraw the entire savings on the MSA for consumption purposes. Instead of that, we specify a target of 10,000€ for the insureds not younger than 65. Every € above the target can then be withdrawn by the insured. This also means that the individual has only to contribute if the balance of the MSA is lower than the target. A similar implementation can be found in Singapore. There, the target is set to roughly 11,000€ (25,000 Sing. $) and the insureds can withdraw from the age of 55 on.

The second MSA plan analyzed shall be the MSA.stano. This policy requires that the insured contributes into the medical savings account up to a specified target which is, however, not age dependent. This plan goes back to Stano’s (1981) proposal of individual health accounts. The medical costs are firstly paid out of the MSA. Beyond that, the remaining medical expenses are covered by the insurer. Stano (1981) suggests a target of 3000$. We, instead, will assume an amount of 5000€ when we simulate because it
Table 4: The Simulated Policies

<table>
<thead>
<tr>
<th>Plan</th>
<th>Contributions</th>
<th>Copayments</th>
<th>Withdrawals for Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>$MSA.ded$</td>
<td>Each year a share of the unused deductible until the age of 65. Then, only if the balance is less than the target.</td>
<td>MSA balance plus the deductible.</td>
<td>With the age of 65, every € above the target.</td>
</tr>
<tr>
<td>$MSA.stano$</td>
<td>Only if the balance is less than the target.</td>
<td>MSA balance.</td>
<td>No withdrawals.</td>
</tr>
<tr>
<td>$MSA.reb$</td>
<td>Each year.</td>
<td>MSA balance.</td>
<td>Every three years, the remaining balance.</td>
</tr>
<tr>
<td>$DI$</td>
<td>No contribution.</td>
<td>The deductible.</td>
<td>No withdrawals.</td>
</tr>
</tbody>
</table>

is found that this improves the performance of such a policy.

A third variation is the $MSA.reb$ which is designed rather similar to a rebate policy. A rebate policy will pay a part of the premium back to the insured if no claims are submitted during a certain time - typically three years. We can implement such policy with the help of a medical savings account. The insured shall be required to pay regularly into the MSA for a period of three years. Every year the full amount of the MSA balance must be used to pay all medical expenses. The insurer covers the remaining costs. After those three years, the insurer can withdraw every € from the MSA and use the amount for consumption. This, of course, requires that the money in the account is not already used up. Likewise the standard rebate plan, the policy holder gets rewarded if she spends the MSA money wisely. The $MSA.reb$, however, is somewhat more transparent than the common rebate policy because the rebate is directly linked with the own contributions into the savings account.

Finally, we will simulate a common deductible insurance ($DI$). The deductible shall be fixed in each year, and the insured has to pay her medical costs up to this amount. The remaining is paid by the insurer if necessary. Figure 8 shows an example how the four policies work where the $MSA.ded$ includes a savings rate of 1. All plans exhibit a deductible or contribution, respectively, of 1000€. Therefore, the premiums $p$ are substantially different.

We already mentioned that the US-type medical savings account is rather
similar to a common deductible insurance in the way that the overall deductible does not depend on the balance of the MSA. Thus, there is no incentive (and no requirement) to deposit more money in the MSA than the amount of the deductible. However, one could also argue that the $\textit{MSA.stano}$ is likewise the American MSA as long as one assumes that the insured is able and willing to contribute enough money to meet the target in each year. The difference between the two points of view lies in the first year payments. With a deductible insurance, the insured has only to pay the accrued medical costs of the first year. But, the policy holder would have to pay the entire amount of the target with a $\textit{MSA.stano}$. In the following years, however, both set-ups are equal in the way that the insured always pays the medical costs up to the amount of the deductible or target, respectively.

**Assumptions and Parameters**

In the following analysis, we need to specify several exogenous parameters. However, to select a certain value is always a somewhat subjective task.

To keep as much similarity to the theoretical model, the preferences and the income in each year of the insureds shall be identical. We will use two different utility functions which exhibit constant absolute risk aversion (CARA) and constant relative risk aversion (CRRA). Because some of the comparative static results of the theory of optimal insurance coverage de-
pend on the form of the risk aversion shown by the insured. The following functional forms are assumed:

\[ u_a(w) = 1 - e^{-\gamma_a w}, \]  

\[ u_r(w) = \begin{cases} 
\ln w & \text{if } \gamma_r = 1 \\
(w^{1-\gamma_r} - 1)/(1 - \gamma_r) & \text{if } \gamma_r \neq 1 
\end{cases} \tag{18} \]

where the subscript \( a \) and \( r \) symbolize CARA and CRRA, respectively. For the parameter \( \gamma_a \), we will take the value 0.0006 which was estimated by Marquis and Holmer (1996). The utility function of (18) is also used by Cutler et al. (1990), where they assume a \( \gamma_r \) of 1 or 10, respectively. Meier (1997) estimates a value of 3 for the constant relative risk aversion. We will use 1 and 3. As annual income, we set the German employee’s real net income of the year 2002 which was roughly 16,000€ (BMG (2003)).

The lifetime utility shall be

\[ v = E \sum_t \rho^{t-1} u_t \]  

where \( \rho \) is the discount factor. We set this to 0.966 according to an estimation of the time preference rate of 3.5% by Meier (1997). Because the savings in the medical savings accounts shall bear interests, we set these to 3.3% \((r = 1.033)\) which equals the average real interest rate of the last 20 years calculated by the Deutsche Bundesbank (2001). Therefore, as in the theoretical model, time preference rate and interest rate are assumed as almost identical.

In order to compute the premium, we have to specify a function which maps the expected indemnity payments by the insurer to the premium to be paid by the insured. According to equation (1) on page 5, this is assumed as

\[ p = p \left\{ \sum_{t=1}^T \{(p - (1 + \lambda)E_i) \times r^{T-t} \} \right\} = 0 \]  

which implies an equal premium in each year. We have not found an estimation for a German or Swiss health insurance loading \( \lambda \). Therefore, we will follow American Academy of Actuaries (1995b) and use a loading factor of 15% \((\lambda = 0.15)\).

\[ \text{For instance, Mossin (1968) shows that decreasing absolute risk aversion is sufficient for a non-negative relationship of the optimal deductible and income. On the other hand, a CRRA utility function and the implied income effect might lead to a reduction of the chosen deductible if the insurer raises the insurance loading. Only a CARA utility can assure that the relationship between insurance loading and optimal deductible is non-negative.} \]
Different copayment levels should also lead to different incurred health care costs due to ex post moral hazard - one can somewhat influence the medical costs after an illness has occurred if one prefers, for instance, to visit more than one physician in order to double check their diagnose. We will address this in two ways: i) We multiply the simulated health care costs by a factor to adjust for the different deductible levels of the Swiss Data. ii) When we assign the specific plans to the individuals in the simulation, we will also take into account the implied copayments where we further distinguish between the source of the money: out-of-pocket or out-of-msa.

The first task is necessary to get the same starting point for all individuals in the sample although they exhibit different copayment levels. Otherwise, the simulated health care cost vary not only due to differences in risk but also because of moral hazard. We take the factors estimated by Werblow and Felder (2002) who use a similar sample than we do. They explicitly try to find the impact of the deductible level on the medical costs due to moral hazard. Their estimated factors are 1 if \( \text{ded} = 0 \), 1.1096 if \( \text{ded} = 1 \), 1.17 if \( \text{ded} = 2 \), 1.887 if \( \text{ded} = 3 \), and 2.2144 if \( \text{ded} = 4 \). Therefore, all the simulated costs get increased if \( \text{ded} > 0 \). Of course, this allows us only to adjust for the size of the medical expenses but not for the frequency of physician visits.

The simulated policies will have different levels of copayments. We, therefore, have to think about how to model their impact on the insured’s medical care utilization. Manning et al. (1987) have shown that such a trade-off exists. We will use the arc price elasticity of demand \( \eta \) estimated by this study of -0.2 to reduce the incurred medical costs according to the current coinsurance level.\(^\text{17}\) Furthermore, we want to distinguish the impact of different sources of copayments on the magnitude of moral hazard. It is believed that a medical savings account might be seen just as another type of insurance and not as savings. Similar to American Academy of Actuaries (1995b), we incorporate this in the way that we assume the individual values her MSA balance to 90% or 10% (\( \psi = 0.9; 0.1 \)), respectively, as real copayments, and that this is also known by the insurer, so that it will compute the premium accordingly.

To keep things tractable, we limit the choice of different parameter constellations to two aspects: the individual’s valuation of the MSA balance as savings or another form of insurance, and the individual’s risk aversion.

\(^\text{17}\) All our simulated policies exhibit a kind of deductible. The coinsurance level of the current year is, therefore, \( \mu = \text{copayments/medical expenses} \). The simulated medical expenses \( m_0 \) are assumed as occurred if \( \mu_0 = 0 \). To set \( \mu_0 = 0 \) would be misleading because the arc elasticity is defined as \( \eta = \frac{\Delta m}{(m_0 + m)^2} \frac{\Delta \mu}{\Delta \mu} \) which results in a factor of \( m = m_0 \frac{1 + \eta}{\mu} \). However, this factor is independent of the increase in \( \mu \). With \( \mu_0 \) slightly higher zero, we get \( m = m_0 \frac{x}{x-\mu} \) where \( x = \frac{\mu + \mu_0}{\mu - \mu_0} \). For instance, an increase of the coinsurance rate from 0.1 to 0.5 together with a price elasticity of -0.2 reduces costs of 1000€ to roughly 765€.
The first one clearly takes effect in one direction: the less the savings are seen as one’s own, the more favorable is a deductible insurance because it reduces the medical costs - and, therefore, the premium - substantially. The individual’s risk aversion should influence the willingness to insure at all: the less risk averse the higher are the benefits of self-insurance. Thus, we distinguish four parameter sets:

<table>
<thead>
<tr>
<th>Set</th>
<th>$\psi$</th>
<th>$\gamma_a$</th>
<th>$\gamma_r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSA High/Insure 90%</td>
<td>0.0006</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>MSA Low/Insure 10%</td>
<td>0.0006</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>MSA High/Self 90%</td>
<td>0.0001</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>MSA Low/Self 10%</td>
<td>0.0001</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Testing the Theoretical Model**

At first, we test the prediction of the theoretical model within the framework of our simulation. We focus on the question whether a positive savings rate of unused copayments will increase the expected utility. In order to do so, we predict for all 25 years old the value of $\pi_i \times m_9i$, and randomly take one individual from each decile of this distribution. For all ten subjects, we simulate their medical expenses up to the age of 100. This is done one hundred times, so that we get for every individual 100 different medical histories. This distribution shall be known to the individual and the insurer, where every path is seen as equal probable. With the above stated assumptions, we can calculate the expected utility and maximize it. To test the prediction, we assign each individual with a deductible insurance at first, and look for the optimal deductible. Then, we slightly increase the savings rate of unused copayments from zero (means common deductible insurance) to 0.1. If the expected utility also increases, we can conclude that a medical savings account is advantageous.

Figure 9 and 10 visualize the results for both utility functions. The picture is rather similar. Obviously, to save every year the entire deductible (savings rate 1, right figure) is almost always welfare decreasing. This can be seen if one compares the left with the right figures. The reason is simply that one would accumulate to much money on the MSA without reducing the premium enough. But, to save for instance only 10% can improve the insured’s well being though only minor. However, both the savings must be seen as one’s own and the risk aversion should be low enough. The first condition is trivial. The second one assures that the individual is willing to accept a minimum of copayments. Only then, the difference is large enough to be recognized. For instance, the average optimal deductible if $\gamma_a = 0.0006$
Figure 9: Differences in $v_a$ Between MSA and Deductible Insurance. Savings Rate= 0.1 and 1

One should not misinterpret the poor performance of a MSA for the individuals drawn from the first and second decile. The first one has medical cost higher than zero but small only in the last year. One could also improve her welfare if one chooses an even smaller savings rate. The second one causes positive expenses not before the age of 65 but then higher than the target. To save cannot decrease the premium and is, therefore, inferior.

All in all, we can conclude that the prediction of the theoretical model is confirmed: to add the option of partly saving unused amounts of a deductible is welfare improving as long as these savings are seen as one’s own money. In
the model, we have implicitly assumed this value to be 1. In the simulation, we have showed that with $\psi = 0.9$ the prediction is still true. However, the more this value shrinks the more becomes a MSA inferior. Unfortunately, we don’t know the real value. It is imaginable that $\psi$ can be influenced by the rules of how to withdraw money out of the savings account for consumption purposes. Therefore, a MSA of the type of $MSA.stano$ or $MSA.reb$ might increase $\psi$ compared to $MSA.ded$ due to the fact that withdrawals or no required contributions are more frequent.

Comparing Individual Insurance Plans

At next, we want to find out which of the four policies is preferred by an expected utility maximizer. We have already established that under the assumption of a large $\psi$, a MSA can be welfare improving because it allows to bear a higher share of the medical costs and, thus, decreases the premium. However, a $MSA.stano$ or $MSA.reb$ are defined in such a way that it is not possible to end up with too large savings. Therefore, they might be even more welfare improving.

Figure 11: Differences in $v_a$ Between MSA.ded, MSA.stano, and MSA.ded to a Deductible Insurance.

We will compare a $MSA.ded$ with a savings rate of 10% and a target of 10,000€ from the age of 65 on. For the $MSA.stano$, we assume a target of 5,000€. The $MSA.reb$ policy shall allow to withdraw every three years. For all policies, we look for the optimal deductible or contribution, respectively. To keep things tractable, we focus on the parameter sets MSA High/Insure and MSA High/Self because we have already showed that a low $\psi$ will unambiguously discriminate against the MSA plans.

At first, we take again the ten individuals drawn from each decile of the prediction of $\pi_i \times m9_i$. This allows us to distinguish between different risk types. Below, we will also do the comparison for the whole sample. Thus,
the first one is more of an individual perspective: For whom it is preferable to use a MSA? Where the latter one allows statements for a community rated insurance.

Figure 11 and 12 show the differences in utilities between the MSAs and a deductible insurance for both parameter sets (left side: MSA High/Insure; right side: MSA High/Self). Again, the most happens if the individuals are willing to accept higher copayments. On the left side of both figures, there is almost no difference perceivable. However, in the case of low risk aversion (right side) both figure speak a clear language: only the policy of type $MSA.ded$ with a savings rate of 10% can improve the welfare of the insureds. The worst is indeed a plan with a rebate policy. The reason is that the savings accumulate in an erratic manner without any link to the fact that the medical costs will almost always be substantially higher in the later years. Over many years, the insured with a $MSA.reb$ saves money without a large effect on the premium. In the same manner, a MSA proposed by Stano (1981) is not directed enough to overcome the disadvantage of periodically saving money without influencing the premium enough. A deductible insurance is then superior because it requires copayments only if they are also used to pay medical expenses in the current year.

As the figures suggest, especially the individual drawn from the tenth decile would gain the most if she starts to save unused copayments. This does not surprise because individuals with high expected medical costs can also decrease their premium substantially if they increase their copayment level. To save the not used deductibles in healthy years allows this in a more dramatic way than a standard deductible insurance. For the individuals from the other deciles, this effect is only minor.

Figure 12: Differences in $v_r$ Between $MSA.ded$, $MSA.stano$, and $MSA.ded$ to a Deductible Insurance.
Comparing Community-Rated Insurance Plans

Let us now turn to a community rated health insurance. Instead of picking out single individuals from all the 25 years old, we now take the whole group and extrapolate their medical expenses up to the age of 100. After that, we compute the average expected utility under different policies, and look for the deductible or contribution, respectively, which maximizes the average welfare. The policies have the same features than those used above. However, this time we explicitly allow to search for the maximizing savings rate of the $MSA.ded$. In addition to the average expected utility of the entire population, we calculate the resulting premium – which shall now be equal for all insureds (community rating). And, we report the overall sum of the lifetime medical costs including the insured’s copayments. The latter gives us a hint which insurance plan has the highest potential to reduce total medical expenditures, and not only the indemnity payments of the insurer, which could also be derived from the size of the premium.

Figure 13: Differences in $v_a$, Premium, and Overall Medical Expenses Between the MSAs and a Deductible Insurance: MSA High/Insure

Figure 13 and 14 show the difference in all three categories to the deductible insurance. The former is computed with the parameter set MSA High/Insure, and the latter with MSA High/Self. For the CRRA utility function, the picture is similar, so that we forgo it. The $MSA.ded$ in Figure 13 exhibit a savings rate of 40%, the other of only 5%.

As one can see, a MSA of the type $MSA.ded$ delivers a higher average expected utility, combined with a lower premium and medical expenses compared to a common deductible insurance. However, this difference is only in the case of the first parameter set relative large; where with the second parameters (lower risk aversion), deductible and $MSA.ded$ are rather similar. Thus, this time the picture has changed compared to the risk av-
justed insurance policies described above: a higher risk aversion increases the advantage of a MSA of type MSA.ded. This is due to the fact that the members of the tenth decile gain the most if everyone switches to a MSA insurance and risk aversion is assumed to be $\gamma_a = 0.0006$, as seen on the left side of Figure 11 on page 28.

**Figure 14:** Differences in $v_a$, Premium, and Overall Medical Expenses Between the MSAs and a Deductible Insurance: MSA High/Self

At least in the first case, a MSA.stano is also able to outperform a deductible insurance. With lower risk aversion, however, this policy is not a good choice. But, compared to a MSA.ded, the difference might be overstated because of the assumed equal $\psi$. As already mentioned above, it is rather reasonable that the savings on a MSA.stano are more perceived as one’s own than with a MSA.ded. This, however, would shrink the advantage of the latter one.

The worst insurance plan is again the rebate policy. The same as stated above applies again: there is no reason to assume that a three year rhythm (or any other short period) of saving and withdrawing can reduce the premium substantially. Even if a MSA.ded implies a low $\psi$, these results would suggest that it is preferable to introduce a MSA as proposed by Stano (1981) instead of a MSA.reb because it seems not to be reasonable to belief that the perception of the savings are very different between a MSA.stano and a MSA.reb.

In the first simulation (MSA High/Insure), the introduction of a MSA.ded would reduce the overall lifetime medical expenses in about 3 Mio.€ compared to a deductible insurance. However, if one assumes a full insurance, this reduction would be almost 24 Mio.€. With a sample size of 1427, these numbers become 2,100€ and 16,700€ per capita. The latter is equal to a 12.6% decrease of the total medical expenses. These numbers, of
course, depend highly on the assumed price elasticity of demand for medical care.

At last, we want to take a look on the implied distribution of copayments among the insureds if one assigns the optimal MSA.ded insurance calculated above (MSA High/Insure). Apparently, each insurance plan which includes a deductible or coinsurance will deviate from a uniform distribution where every insured has no copayments at all. The simulation allows us to compute a cumulative distribution of the lifetime copayments; this is shown on the left side of Figure 15. The optimal MSA includes a deductible of 1,300€ and a savings rate of 40%. As one can see, this implies a considerable amount of copayments. Only 10% of the insureds have paid less than 22,000€ during their lifetime. The median lies around 44,000€. With a calculated premium of 870€ per year, this means that 50% of the insureds paid roughly the same amount of premium and copayments.

**Figure 15:** Cumulative Distribution of the Lifetime Copayments and the MSA Balance When 64 Years Old

On the right side of Figure 15, the distribution of the MSA balance among the 64 years old is drawn. This is the year before the target takes effect so that we can expect that a major part of the money above the target of 10,000€ can be withdrawn in the next year. Assuming our optimal MSA.ded policy, in about 23% of all the insureds could expect some money to be withdrawn. 77% have a MSA balance below the target. In addition, there is a small group (10% of the sample) who would gain almost 10,000€ (19,815.42 − 10,000).

Compared to a full insurance, these numbers imply an enormous unequal distribution among the insureds. However, everyone who pays during her lifetime less than 60,000€ of copayments would still gain from a switch to a MSA plan. The reason is that with a full insurance, the premium would increase from 870 to almost 1900€ per year. The individuals are insured
in about 57.5 years on average until they die. Therefore, a MSA.ded would save them over 57 years an amount of 1,030€ per year. Thus, as one can see in Figure 15, in about 90% of the insureds would be better off.

4 Conclusions

The purpose of this paper was to compare a common deductible insurance with medical savings accounts. The theoretical model has shown that a MSA can reduce the insured’s income risk: a risk averse individual would, therefore, prefer a medical savings account. In addition, we have proved that a situation where the insureds can expect increasing medical expenses in later periods, a MSA is superior to a common deductible insurance because it reduces the premium substantially.

To overcome the limitations of a pure theoretical analysis, we have employed a simulation which has allowed us to test not only the predictions of the theoretical model, but also to go further and to analyze a MSA proposed by Stano (1981) and a rebate policy. In addition, we could incorporate aspects of moral hazard and even distinguish between the impact of the copayments on the demand for health care if the money is paid out-of-pocket or out-of-msa.

The data of the Swiss insurance company CSS has shown that the medical expenses of their insureds are strongly increasing in age. Therefore, a MSA should perform well according to the theoretical model. However, as the simulation has shown, this is true only if the savings rate is rather small. Furthermore, one has to assume that the individuals value the money on the MSA as their own savings and not as another type of insurance. The impact of the form of risk aversion is negligible.

A MSA.ded is compared to a MSA with a specified target (MSA.stano) and a rebate policy superior as long as the valuation of the savings do not substantially differ. In particular, a rebate policy implemented as a type of MSA seems to be a rather bad choice. This is remarkable because of the relative widespread use of similar plans among the private health insurer. Of course, the real world plans differ from the MSA.reb considered here in the way that they do not feature a medical savings account. Instead, the insured can get a premium reduction if she claims no reimbursement of medical costs during a specified period. However, as long as the premium reduction does not increase with more and more claim free years, the static nature of this plan is inferior compared to a medical savings account. The reason is that the copayments18 are not linked with the expected increase of medical expenses in older years.

A switch from a community rated full insurance to an insurance with

---

18 The copayments are indirectly the higher premiums if one sets up a rebate policy compared to an insurance without rebate payments.
a medical savings account and a moderate savings rate could increase the overall welfare among the insureds of a risk pool. However, one would have to assume that the higher level of copayments would also decrease the overall medical expenditures due to reduced moral hazard. Otherwise, the implied redistribution of income from the bad risks to the good risk could not be offset by a large enough premium reduction.

References


A The Proofs

A.1 Proof of Lemma 2

With $\delta = 0$ and $v_d' \geq 0$

$$-p_d' v \geq \mathbb{E}_{hh}[ru_2'] + \mathbb{E}_{hl}[u_1'] + \mathbb{E}_{lh}[1 + ru_2'], \quad (21)$$

where

$$-p_d' = \frac{1 + \lambda}{1 + r} (\mathbb{E}_{lh}[1] + \mathbb{E}_{hl}[r] + \mathbb{E}_{hh}[1 + r]). \quad (22)$$

Setting (22) into (21), multiplying both sides by $r$ and dividing by the term in the brackets of (22), we get

$$r \frac{1 + \lambda}{1 + r} v' \geq \frac{\mathbb{E}_{lh} ru_2' + \mathbb{E}_{hl} ru_1' + \mathbb{E}_{lh} [ru_1' + ru_2']}{\mathbb{E}_{lh} 1 + \mathbb{E}_{hl} r + \mathbb{E}_{hh} (1 + r)}. \quad (23)$$

Because of $u'_{1,2} = u'(y - p - d)$ for every $u'$ on the rhs of (23), we can place it outside the brackets. Simplifying yields the result of the lemma.

A.2 Proof of Corollary 1.1.

At the point $\delta = 0$ the arguments of the utility function are identical in both periods: income minus premium and medical costs. If the medical costs are independent and identical distributed, the difference between the expected marginal utilities without the multiplier $(d - a_1)$ becomes zero. However, the multiplier $(d - a_1)$ is high exactly in those situations where the medical costs are low in the first period. Therefore, $(d - a_1)$ weights those cases more where the difference of the marginal utilities is negative.

Proof. With $rp = 1$, $\mathbb{E}_{gg}((d - a_1)(u_1' - u_2'))$ corresponds at the point $\delta = 0$ to:

$$\int_0^d \int_0^d (d - a_1)(u_1' - u_2') f(a_1)f(a_2) da_2 da_1 \quad (24)$$

with $u_1' = u(Y - P - a_1)$ and $u_2' = u(Y - P - a_2)$. Cutting the intervals $[0, d]$ of $a_1$
and \( a_2 \) into \([0, d/2]\) and \((d/2, d]\) leads to
\[
\int_0^{d/2} \int_0^{d/2} (d - a_1)(u'_1 - u'_2)f(a_1)f(a_2)da_2da_1 \\
+ \int_{d/2}^{d} \int_{d/2}^{d} (d - a_1)(u'_1 - u'_2)f(a_1)f(a_2)da_2da_1 \\
+ \int_0^{d/2} \int_{d/2}^{d} (d - a_1)(u'_1 - u'_2)f(a_1)f(a_2)da_2da_1 \\
+ \int_{d/2}^{d} \int_0^{d/2} (d - a_1)(u'_1 - u'_2)f(a_1)f(a_2)da_2da_1. 
\tag{25}
\]

The signs of the first two summands are unknown until now. However, one can see that the third summand is negative and the fourth positive because of the bisection. Furthermore, the assumption of independent and identical distributed medical costs ensures that the following applies:
\[
\left| \int_0^{d/2} \int_{d/2}^{d} (u'_1 - u'_2)f(a_1)f(a_2)da_2da_1 \right| = \\
\left| \int_{d/2}^{d} \int_0^{d/2} (u'_1 - u'_2)f(a_1)f(a_2)da_2da_1 \right|. 
\tag{26}
\]

Multiplying both sides of (26) with \((d - \frac{d}{2})\) and recognizing that
\[
\left| \int_0^{d/2} \int_{d/2}^{d} (d - a_1)(u'_1 - u'_2)f(a_1)f(a_2)da_2da_1 \right| \\
> \left| \int_0^{d/2} \int_{d/2}^{d} (d - \frac{d}{2})(u'_1 - u'_2)f(a_1)f(a_2)da_2da_1 \right| 
\tag{27}
\]

and
\[
\left| \int_{d/2}^{d} \int_0^{d/2} (d - a_1)(u'_1 - u'_2)f(a_1)f(a_2)da_2da_1 \right| \\
< \left| \int_{d/2}^{d} \int_0^{d/2} (d - \frac{d}{2})(u'_1 - u'_2)f(a_1)f(a_2)da_2da_1 \right| 
\tag{28}
\]

applies, leads to the conclusion that the absolute value of the third summand in (25) is higher than the value of the forth summand. Therefore, the sum of both is negative. The first two summands could be divided accordingly, and the argumentation is the same as before. Again, one gets a sum of two terms which is negative and two terms with unknown sign. This halving could be done again and again, so that the overall sum is negative. Therefore, the condition for a positive rate of saving from lemma 3 is met.

A.3 Proof of Proposition 2

Define \( \mathcal{D} = \{(d, \delta) : V'_1 \geq 0 \land \delta > 0\} \). For all \((d, \delta) \in \mathcal{D}\) equation (23) becomes
\[
\frac{r}{1 + r'} + \frac{1}{1 + r'} \delta \left[ \mathbb{E}_{l}(u'_1 - u'_2) + \mathbb{E}_{h}(u'_1 - u'_2) \right] + u'_d. 
\tag{29}
\]
Setting (29) into (10) leads to

\[
v' \delta \geq -E_{ll}[(d - a_1)(u'_1 - u'_2)] + E_{lh}\left\{(d - a_1) \left[ r \delta \{E_{ll}[u'_1 - u'_2] + E_{lh}[u'_1 - u'_d]\} + u'_d - u'_1 \right]\right\}. \tag{30}
\]

Therefore, to get \(v' \delta > 0\) for all \((d, \delta) \in D\) we need that

i) the second summand of (30) is positive,

ii) and the rhs of (30) is positive, too.

If i) holds, then, clearly, ii) can be ever established by increasing \(\phi \equiv F(lh)/F(ll)\). Let \(\phi_1\) be the lowest \(\phi\) where rhs is positive for all \((d, \delta) \in D\). So, we only have to show that i) is true.

The second summand of (30) is positive whenever

\[
r \delta \{E_{ll}[u'_1 - u'_2] + E_{lh}[u'_d - u'_1]\} < (u'_d - u'_1) \{E_{lh}(1 + r \delta) + E_{hl}r + E_{hh}(1 + r)\} \tag{31}
\]

On the lhs of (31), the first summand in the brackets should be negative because of the transfer of money from period one to two. Then, due to \(E_{lh}[u'_d - u'_1] \leq (u'_d - u'_1)\), condition (31) is satisfied whenever

\[
r \delta < E_{lh}(1 + r \delta) + E_{hl}r + E_{hh}(1 + r). \tag{32}
\]

However, if \(E_{ll}[u'_2 - u'_1] > 0\), we have to increase the required \(\phi\) until \(E_{ll}[u'_2 - u'_1] + E_{lh}[u'_d - u'_1] \leq (u'_d - u'_1)\); and condition (32) can be applied. Let \(\phi_2\) be the lowest \(\phi\) where this is true for all \((d, \delta) \in D\).

Substituting 1 by \(E_{ll}1 + E_{lh}1 + E_{hl}1 + E_{hh}1\) in (32) leads to

\[
E_{ll}[r \delta] - E_{ll}[1] < E_{hh}[r(1 - \delta)] + E_{hl}[r(1 - \delta) + 1]. \tag{33}
\]

The rhs of (33) is non-negative. Let \(\phi_3\) be the lowest \(\phi\) where the lhs of (33) becomes negative for all \((d, \delta) \in D\).

To sum up, \(v' \delta > 0\) for all \((d, \delta) \in D\) if the joint distribution function exhibits a \(\phi \geq \max\{\phi_1, \phi_2, \phi_3\}\).