Does private information affect the insurance risk?

Evidence from the automobile insurance market.

Sara Arvidsson

VTI*/CTS**

Abstract: This paper empirically investigates the effect of policyholders' private information of risky traffic behavior on automobile insurance coverage and ex post risk. It combines insurance company information with private information data that is not accessible to the insurance company and shows that being unable to reject the null of zero correlation is not necessarily consistent with symmetric information in the automobile insurance market. The results are twofold: In contrast to much of the previous work, a positive significant correlation for three groups of policyholders is found, consistent with the adverse selection prediction. Besides, private information about risky traffic behavior increases ex post risk while it both increases and decreases the demand for extensive insurance which supports the hypothesis that adverse and propitious are present simultaneously in this market.

*Swedish National Road and Transport Research Institute (VTI)
Box 920, 781 29 Borlänge.
sara.arvidsson@vti.se
**Centre for Transport Studies, Royal Institute of Technology
Teknikingen 78 B, 100 44 Stockholm.

Acknowledgement: I wish to thank Länsförsäkringar AB for the data; Lage Niemann and Björn Johansson for helpful discussions about the data, the company and its market. Thanks to Jan-Eric Nilsson (VTI), Lars Hultkrantz (Örebro University), Daniela Andrén (Örebro University), Henrik Andersson (Toulouse School of Economics) and participants at seminars at Örebro University and VTI for useful comments on previous drafts of this paper.

1. Introduction

Asymmetric information has for long been alleged to cause inefficiencies in insurance markets. However, the empirical findings regarding the automobile insurance markets have been ambiguous as to whether or not to support the core prediction that individuals with extensive coverage are more likely to be high risks for the insurer. The majority of previous work has interpreted the lack of a significant coverage-risk correlation as absence of contract-relevant information asymmetry. Explanations such as absence of useful private information and policyholder inability to act on private information have been used. Since adverse selection theory impregnates many areas, primarily relying on the theoretical prediction, it has important implications for policy decisions. Empirical research in this area is therefore highly relevant, not only to the economist. Cohen and Siegelman (2010) argue that rather than trying to resolve the question of the existence of adverse selection once and for all, future work should try to identify circumstances under which one may expect to find evidence of relevant information asymmetry. Since market heterogeneity may play an important role, we believe it is difficult to generalize across insurance markets and between countries. It is furthermore reasonable that the correlation structure differs across subsets of policyholders.

This paper seeks to contribute to the empirical risk-coverage literature by testing information asymmetries in a less generalized setting. Our work differs in three major ways: first, we include policyholders' private information about risk: second, we use several subgroups that correspond to the insurer's classification: third, we consider specific characteristics in the market that may affect the outcome. Our hypothesis is that both adverse and propitious selection is present simultaneously in the automobile insurance market. If so, the correlation structure may differ from what we expect according to the classical adverse selection prediction of a positive coverage-risk correlation. In combination with a negative coverage-risk correlation, predicted by propitious selection, the correlation effects may cancel out, producing zero correlation. This implies that inefficiencies may exist in the market in spite of the standard correlation structure predicted by theory.

We use a rich data set of automobile insurance policies, provided by one of Sweden's largest insurance companies. The core difference to previous work on automobile insurance data is that we add data on the policyholders' private information of risky traffic behavior, which is inaccessible to the insurer. Private information is represented by observed traffic safety violations in terms of on-the-spot-fines and convictions for traffic offences.[1] The advantage of this data is that we are able to directly observe the effect of private information in this particular market, which implies that our conclusions are not all dependent on the existence of a risk-coverage correlation. Like Cohen (2005), we use a sample of new policyholders, since the information asymmetry between insurer and policyholders is likely to be largest in the beginning, but analyze more homogenous subgroups than previously done. Furthermore, we put a restriction on vehicle age since it may be an important determinant of choice of coverage and how the vehicle is used. Conditional on a close replication of the risk classification, made possible by access to the insurers actuarial predicted risk classification, we test whether the existence of private information confirms the prediction of adverse and/or propitious selection theory.

First, we use the correlation test, as suggested by Chiappori & Salanié (2000), where we allow for propitious selection. We also test the null hypothesis of zero correlation that neither adverse nor propitious selection dominates. If there exists a significant correlation between risk and coverage, the null is rejected. Second, we use an approach suggested by Finkelstein and McGarry (2006), where we directly observe the effect of private information.

The results indicate the existence of residual private information that predicts the risk. This residual private information is positively correlated with insurance coverage for three groups; females in age group 18-21, females in age group 30-39, and policyholders of both sexes in age group 50+. Further, our results indicate that the policyholders' private information about traffic offences and convictions is positively related to cases where the policyholder was fully or partially at fault in the reported claim. This implies that the policyholders have information, unobservable to the insurer, that predicts the ex post risk. Private information about speeding is mainly positively related to having extensive coverage, which confirms the adverse selection prediction. Private information about convictions is essentially negatively related to extensive coverage, which in turn implies that the insurer is left with a propitious selection. This pattern remains consistent both where the correlation test suggests adverse selection and where the null of symmetric information cannot be rejected. Previous research has established that violations have a significant effect on crash rate and that the risk acceptance between different violations are disparate. This difference can be attributed to the social norm that

[1] Note that private information may also be related to good risks, i.e. absence of convictions or on-the-spot fines.

speeding is more accepted compared to other traffic safety violations (see Åberg; 1998 for a review). Our observed difference of the effects of traffic safety violations on the demand for insurance may also mirror this social norm.

The rest of the paper is organized as follows. Section 2 provides a summary of prior theoretical and empirical research with a focus on insurance markets. The section also contains information about the insurance coverage and risk classification in the Swedish automobile insurance market. Section 3 describes the empirical approach in terms of data and econometrics in more detail. Section 4 presents the results and Section 5 concludes the paper.

2. Background

A. Previous work

Ever since the 1970s the theoretical research regarding asymmetric information has developed at a quick pace. The prediction is that asymmetric information is a fundamental problem in most insurance markets: Policyholders are heterogeneous in risk and this risk level is private (hidden) information that is important for the contract but unobservable to the insurer. According to the standard interpretation, the asymmetry results in a situation where high risk individuals buy extensive insurance coverage. This predicts a positive correlation between ex post risk and extensive coverage and implies that those with insurance constitute an adverse (bad) selection of risks (Rotschild and Stiglitz 1976; Akerlof 1970; Bolton & Dewatripont 2005 & Salanié 2005). In addition, the insured may undertake private (hidden) actions that affect the risk and thereby the contract. An individual with insurance is then less cautious since s/he does not fully carry the financial risk of an accident. This is known as moral hazard. Both adverse selection and moral hazard both produce a positive correlation. Disentangling them empirically is generally viewed as difficult and is beyond the scope of this paper.

Several studies, both theoretical and empirical, have suggested the possibility of propitious (favorable) selection. Policyholders are heterogeneous not only in their probability of loss (as in the adverse selection model) but also in their aversion to risk. Along the same line of reasoning, the policyholder may perform preventive actions that reduce the risk in the contract. These individuals have a high demand for insurance and are good risks ex post.

From the perspective of the insurance company, these types represent a propitious selection of risks (Hemenway 1990, DeMeza & Webb 2001; Finkelstein & McGarry 2006; Fang & Silverman 2006). DeDonder and Hindriks (2009), however, show that, under some mild regularity assumptions, this prediction still does not imply a negative correlation between risk and insurance coverage in equilibrium. The reason is that there is a moral hazard effect: after obtaining insurance the policyholder becomes less risk averse since most of the economic risk is transferred to the insurer.

Empirical research regarding asymmetric information has lagged behind and did not significantly evolve until the 1990s. As discussed by Chiappori and Salanié (1997), data from insurers is well suited for studies of asymmetric information, because it records choice of coverage and outcome (claim or not), as well as many characteristics of the policyholders. Empirical studies have used data from different insurance markets and found evidence of a coverage-risk correlation (See for example Cutler; 2000 and Finkelstein and Poterba; 2004).

Still, empirical tests on property/liability insurance, where automobile insurance data has been used, do not provide any strong evidence of information asymmetries that affect the level of risk in the contract (see Chiappori and Salanié (2003) for a review). Three early studies suggested the presence of a positive correlation, but these were later criticized as unreliable. The first and second, Dahlby (1983, 1992) found evidence in favor of adverse selection in the Canadian automobile market, but these studies did not have information on individual coverage. The third, of Puelz and Snow (1994), used data on individual policies from the US automobile market. Their result has since been questioned, one reason being that they did not have information about some of the variables affecting risk type that the insurer had. That is, they applied their analysis to an insufficient information set, which may have resulted in a spurious correlation driven by omitted variables. Dionne, Gouriéroux and Vanasse (2001) do not find any evidence of information asymmetries using French automobile insurance data. They suggest that the insurers' information set is sufficient if non-linear effects, not considered by Puelz and Snow, are taken into account. A sufficient risk classification implies that there is no residual adverse selection in each risk class, since groups are homogenous in risk.

Although these studies have built a bridge between theory and practice, the findings are not consistent with the theoretical predictions in the insurance market. To overcome previous difficulties, Chiappori and Salanié (2000) (hereafter C&S) suggest a simple and general test of the presence of asymmetric information. Using French individual data covering one year (1989) with information on 1 120 000 contracts and 120 000 accidents, they focused on a subset of 20 716 drivers with less than three years of driving experience.[2] This group was assumed to consist mainly of young drivers.[3] To test the adverse selection prediction they suggested a correlation test between coverage and ex post risk, and they concluded that the market did not suffer from information asymmetries since they could not reject the null of symmetric information.

Cohen (2005) argues that young drivers may not have private information since they have not learned their own risk type. The hypothesis is that there is a learning effect involved; when the policyholders learn their risk type they develop private information. The study takes several implications of the previous critique into account and uses a rich data set of the first five years of one start up insurer in Israel. The data covers 216 524 policies where a subset of new policyholders with 104 639 policies is used in the analysis. When applying the C&S correlation test on policyholders with less than three years of driving experience, the results are confirmed since no significant correlation is found. However, for a group with more than three years of driving experience, Cohen finds a significant negative correlation that rejects the null of symmetric information. The main conclusion, as drawn from results that indicate that low deductible contracts are associated with more claims, is that the market is characterized by the positive correlation predicted by the classical adverse selection theory.

Cohen and Einav (2007), using Israeli automobile insurance data, provide evidence that, conditional on observables, risk and risk aversion are positively correlated (0.86). Their conclusion is that such a correlation makes it even more likely to find evidence of adverse selection in the automobile insurance market. They argue that risk in this market differs compared to other markets. Taking precautions, like driving slow or (too) carefully, may expose the policyholder to greater risk.[4] They furthermore argue that the correlation

[2] Data was provided by the French federation of insurers (FFSA), which groups 21 companies and constitutes 70 percent of the automobile market. In 1990 they conducted a survey of its members. The sampling rate was 1/20 and the resulting data included 41 variables for 1 120 000 contracts and 25 variables for 120 000 claims.
[3] One reason why they focus on a young sample is that they believe the heteroskedasticity problem is less severe than in a sample with a mixture of more senior drivers.
[4] With regards to speed distribution, Solomon (1964) showed that most accidents on main rural highways involve drivers who are either driving much faster or much slower than the mean traffic speed. This means that the relationship between accident involvement rate and the deviation from the mean traffic speed is U-shaped.

coefficient may be highly sensitive to what measure of risk and risk aversion one is using since there may be omitted factors that may be related to both dimensions.[5] The policy analyzed does not cover at-fault accidents. However, it may be interesting to separate out this category of claims, since a risk-averse individual may report accidents where s/he was not at all to blame. This implies that a measure that considers a wider range of claims may not truly reflect the level of risk of the policyholder, which can affect the correlation between risk and risk aversion. Hence, claims where the policyholder was at fault, as studied in this paper, may not have a correlation structure similar to the one found by Cohen and Einav.[6]

Finkelstein and McGarry (2006) consider the policyholder's private information about risk in the long-term medical care insurance market. They examine the effect of the policyholders' private beliefs of their chances of ending up in long-term medical care in the next five years. This information is unobserved by the insurer. Their findings indicate that two types of individuals buy insurance; those with private beliefs that they are high risks and those with a strong taste for insurance. Ex post the former is a higher risk and the latter a lower risk to the insurer. One explanation of the inability to reject the null of zero correlation may therefore be that several risk types demand more insurance. In such a market the correlation structure can look different from what we expect according to theory. An absence of significant correlation might therefore not imply absence of asymmetric information relevant for the risk in the contract.

B. Social norms and traffic safety violations

Research by psychologists has been able to demonstrate that road crashes are largely attributable to driving violations, such as drunken driving and speeding studied in this paper (see Forward 2008 for a review). Åberg (1998) provides a research summary showing that rules and regulations may increase traffic safety by changing behavior. The best predictor of behavior is the intention to behave in a certain way, which is determined by both attitudes and

---

However, according to the review by Aarts and van Schagen (2005) none of the relatively new studies show that vehicles that move (much) slower than the surrounding traffic has an increased crash rate.

[5] They provide examples like the intensity of vehicle use; risk-averse individuals may be more exposed to accident risk because they drive more per year, which could explain the positive correlation.

[6] They find that the individuals classified as "good driver" by the insurer have a lower risk, while they appear to have lower risk aversion. The exact functional form for the classification good driver is unknown.

social norms.[7] Åberg (1993) concluded that attitudes and social norms were important for the decision of a sample of Swedish male drivers to drive after alcohol consumption.[8] Åberg and Rimmö (1998) survey drivers' self-reported behavior and find that drinking and driving was the violation that was reported least frequently, while speeding was reported as the most frequent violation. Forward (2006) reports that drivers usually find speeding acceptable. Further, Forward et al (2000) find that immigrants are less inclined to exceed speed limits than Swedish residents. The longer the respondents lived in Sweden, the more likely they were to exceed speed limits. The general opinion of immigrant respondents in this study is that there are fewer drunken drivers in Sweden than in their home country. Even though violations have a significant effect on crash rates the risk acceptance between violations differs, which may be attributed to the social norm. Furthermore, Guppy (1993) found that British drivers with prior convictions for speeding or drunk driving in general perceived themselves as less likely to have an accident compared to individuals with no offences.

C. Automobile Insurance and premium pricing in Sweden

Swedish law requires all vehicle owners to have a Traffic Insurance, which is a liability insurance that covers accident damage inflicted to other drivers and their cars. This is the minimum possible coverage offered. In addition the insurance companies offer Limited Damage Insurance and All Risk Insurance, the later being the most extensive coverage on offer since it also indemnifies damages to the insured's own car when the policyholder is at fault in the claim. All Risk Insurance is typically differentiated by the value of the deductible, our particular insurer offering a lower (3000 SEK) and a higher deductible (5000 SEK). Thus, All Risk Insurance with the lower deductible provides the most extensive coverage. It is also possible to purchase a complementary coverage called Additional Insurance, which provides extra service such as a replacement car if something happens to the insured car. The most typical comprehensive coverage in Sweden is All Risk Insurance, which we focus on in this paper.

Swedish automobile insurance companies base their premium classification on three main categories: risk characteristics related to the driver, the vehicle and the residential area. To

[7] Jöreskog and Sörbom (1989) show that attitudes are important for drunken driving decisions and that social norms are correlated with attitudes.
[8] The study shows that Swedes in general had extremely negative attitudes to drunken driving.

establish pricing, information that statistically affects the expected cost of offering insurance is used. In this way insurers develop a risk classification that is associated with observable characteristics. The insurance contracts are thereafter divided into homogenous groups of risk according to observable characteristics. Individuals in the same group are charged the same insurance premium since they are considered homogenous in risk. Swedish insurers use their own formula for determining insurance premiums.

The insurers do not share information about previous claims, so the market structure is similar in that respect to the Israeli market studied by Cohen (2005). Besides, some pricing variables are based on the policyholders' self reports, such as residential area, previous claims, and the owner (main user) of the vehicle. This implies that policyholders generally have incentives to report untruthfully to receive a lower premium. A latent threat, though, is the reduced indemnity the policyholder may receive if an untruthful reporting is detected. This threat may not be credible to the policyholder since the possibilities for insurance companies to prove this opportunistic behavior are limited.[9] These are examples of practical consequences of asymmetric information that obstruct the construction of homogenous groups and thereby premium pricing. Such practical consequences also imply that adverse and propitious selection can be present simultaneously within groups that are considered homogenous according to the risk classification.

3. The empirical framework

To investigate the nature of private information we use a rich data set that contains both the individuals' (partially) observable traffic risk and the ex post risk. Since automobile insurance is a property/liability insurance the contracts rather than the policyholders are considered.[10] The insurer makes three main assumptions regarding the contracts. First, there is independence between contracts, the outcome for different insurance policies being independent. Second, there is time independence in that the outcomes in two separate time intervals are independent. Third, homogeneity is assumed: an outcome with the same

[9] The incentive to report untruthfully is equivalent to theoretical non-binding incentive compatibility constraints.
[10] Note that a policyholder may have several contracts that are viewed as different risks by the insurer. An example is a policyholder who insures vehicles of different brands.

exposure has the same distribution within a risk group. [11] We therefore consider a repeated contract as a new observation and do not consider dependency between periods and between contracts owned by the same individual.[12]

A. Data

The automobile insurance data used in this study comes from an automobile insurance provider in Sweden with 24 regional subsidiaries located in all the counties in Sweden; its market share is approximately 30 percent of the property insurance market. The data set contains information about 2 424 525 insurance policies and 584 425 claims and covers three years (2006-2008). Most of the contracts are repeated and the number of observations when including those are 9 342 749. Each observation includes all the information that the insurer has about the policyholder, vehicle and contract characteristics.

We also add data on the policyholders, which we can access as researchers but is not available to the insurer. This data represents the policyholder's private information about risky driving behavior. Data on the number of convictions for traffic safety violations are registered by the Swedish National Council for Crime Prevention (BRÅ). These are cases where sanctions are carried out by an attorney. Examples include convictions for driving while intoxicated and driving carelessly: that is, traffic violations that lead to more serious sanctions than on-the-spot-fines. Data on on-the-spot fines comes from the RIOB register of the Swedish National Police Board (RPS). The fines are divided into speeding and other traffic offences such as running red lights, overtaking at crossings, and other offences due to risky behavior or vehicle flaws. Since RIOB is cleared periodically, it is possible to receive at most five years from the current year.

[11] There exist several examples when these conditions are violated. One example, already discussed, is untruthful reports of the policyholders, which violates homogeneity. Furthermore, if two vehicles insured by the same insurer are involved in a collision with each other, the independence between contracts could be violated.

[12] An individual (or contract) may appear as several observations if he or she owns several cars, make more than one claim or if any changes is made in the contract. About 25% of our sample of new policyholders (363 158) appear as two or more observations, and 11% have three observations or more. We performed a sensitivity analysis of dependency between individuals in unreported regressions. First, including only the first observation (first contract) of the individual and second, we cluster adjusted the standard errors with respect to the policyholder-id and the results seem robust. In this paper we however consider the insurance contracts rather than the policyholder and 22% of the contracts appear as two or more observations, 5% as two observations or more. Note that if a change is made the contract will have a new duration and is thus a repeated contract. This implies that the only time a contract with the same date will appear as more than one observation is when more than one accident occurs (approximately 0.7%). We therefore not consider dependency between time periods, contracts and individuals.

Data in respect of on-the-spot fines and convictions has been merged with the insurance and claim files by BRÅ for our project. Finally we have merged the insurance and claim files and cleaned the data. Each observation includes the following information:

1. Demographic characteristics of the policyholder: individual id-number, year of birth, gender, home district and self-reported number of kilometers driven per year.
2. Residential area risk classification: the actuarial predicted risk in the neighborhood where the policyholder lives. Each type of insurance coverage (Traffic Insurance, Limited Damage Insurance and All Risk Insurance) has a classification. All policyholder has each classification regardless of coverage.
3. Car characteristics: vehicle model, brand, construction year, size of engine and vehicle-id.
4. Vehicle risk classification: the actuarial risk classification regarding the vehicle. As with residential area risk classification, each type of insurance coverage has a risk classification regarding the vehicle.
5. Private information: The number of on-the-spot fines for speeding or other traffic offences of the policyholder during 2004-2007, and the number of convictions a policyholder had during 1973-2007.
6. The type of policy purchased: Traffic Insurance (required if the car is in use but not if it is deregistered), Limited Damage Insurance, All Risk Insurance (not generally required for new cars since most manufacturers provide insurance) and Additional insurance.
7. Deductible Choice: The only contract providing deductible choice (high or low deductible) is All Risk Insurance.
8. Premium: The price of the insurance policy.
9. Period covered: From date and to date for each period in the contracts. The number of days with insurance is 1-365 days during one period.
10. Realization of risk: Claims submitted by the policyholder and information on which insurance covers the claim. It is also possible to identify the level of at-fault in the claim (none, partial or fully responsible).
11. Driver information: The insurer's information on the identity of the reported driver in an accident (not necessarily the policyholder), age, gender and personal identity number and private information according to (5). Note that additional drivers are

private information to the policyholder since the premium is not dependent on drivers other than the vehicle owner.

12. Other variables: Household identity, two or more policyholders in the same household share the same household-id.

B. Descriptive Statistics

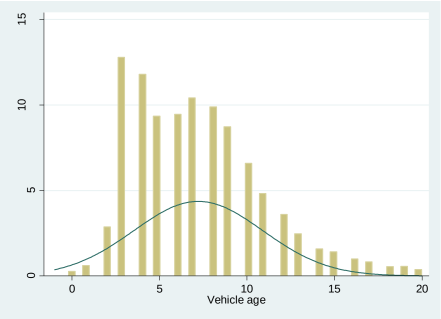
As with Cohen (2005) our focus is on new customers. The reason is that the information asymmetry is likely to be larger between the insurer and new policyholders than for long-term customers who may have demonstrated their type to the insurer. We further divide the policyholders into homogenous age and gender groups that correspond to the actuarial model used during 2006-2008. This provides us with ten groups on which we perform the analysis.

We consider coverage and ex post risk for individuals who joined the insurer in 2007 and 2008 and include all contracts signed by new policyholders in 2007 and observe these contracts until they expire. For new policyholders in 2008, we observe all contracts signed in 2008 until they expire or until the end of 2008 when data was collected. This implies that data is censored for 2008 since we cannot observe the outcome in all contracts.[13]

We restrict our analysis to vehicles of age 3-20. The restriction on vehicle age is due to new vehicles generally having a motor vehicle damage warranty that corresponds to All Risk Insurance. This affects the choice of purchasing more extensive coverage.[14] We also expect that All Risk Insurance is less likely for older vehicles due to a lower economic value. As can be seen in Figure 1 the data confirms that the number of vehicles with All Risk Insurance increases when the vehicle is three years old and decreases as the vehicle gets older. We also perform a sensitivity analysis on this restriction.

[13] We performed a sensitivity analysis of the correlation test by using only new policyholders 2007 for whom we observe the whole lifespan of the contracts, the results can be found in Table 1 in Appendix B.
[14] Approximately 15 percent of the policyholders tend to have All Risk Insurance on vehicles below three years of age. One reason is that the deductible for the warranty is very high for some vehicle makes and some brands do not come with a warranty.

Figure 1. All Risk Insurance and vehicle age.



Note: Vehicle age is -1 to 20: a negative age is possible in cases where the policyholder owns a vehicle of the latest vehicle year model.

Table 1 provides some descriptive statistics of some of the variables for the whole sample and the subset of all new policyholders 2007 and 2008.

Table 1. Descriptives of the whole sample and new policyholders.

	Whole sample	Min	Max	New policyholders	Min	Max
Number of contracts	9 342 749			363 158		
All Risk Insurance with low deductible	44%			33%		
Average year of birth	1955	1898	2004	1964	1908	1997
One conviction	7.1%			6.5%		
Several convictions	3.3%			4.3%		
Total number of convictions		0	136		0	117
Traffic offences	11.4%	0	38	10%	0	7
Speeding tickets	7.2%	0	8	12.7%	0	30

In general young individuals have a higher share of on-the-spot-fines for traffic offences compared to the older groups. This indicates that younger individuals are riskier.[15] On the other hand older individuals have a higher share of convictions compared to young individuals. This is likely explained by seniority since higher exposure increases the probability of being observed, and convicted, for a traffic safety violation. Furthermore, males constitute a higher share of vehicle owners compared to women, and women tend to have a

[15] Note that fines come from RIOB with time period 2004-2007, while convictions come from BRÅ and time period 1973-2007. This implies that it is likely that younger groups have a higher share of fines compared to the older groups. The probability of having one or several convictions, however, increases with age.

lower share of convictions and at fault claims. This implies that women and young individuals have lower frequencies, especially for convictions, compared to males.

C. Econometrical approach

Our first step in the data analysis is to examine the relationship between insurance coverage and ex post risk where the policyholder is held fully and partially responsible in the reported claim. Since the purpose is to investigate if risky traffic behavior tends to affect the probability of risk ex post, at-fault is an informative indicator. One reason is that a pure "bad luck" accident is not as likely to be affected by the policyholder's risk type as at-fault claims.

We first apply the bivariate probit model suggested by C&S to test for adverse or propitious selection.

$$c_i=1(X\beta_1+\varepsilon_i>0) \qquad (1)$$

$$y_i=1(X\delta_1+\eta_i>0) \qquad (2)$$

(i = contract)

The dependent variable of equation (1) represents the choice of a particular contract, $c_i = 1$ if the policyholder has the highest possible coverage, that is, All Risk Insurance with low deductible (3000 SEK) and $c_i = 0$ if less coverage is bought (All Risk Insurance with high deductible (5000 SEK), Limited Damage Insurance or Traffic Insurance).

The dependent variable of equation (2) represents the occurrence of an at-fault claim, $y_i = 1$ if the policyholder reported a claim where s/he was partially or fully responsible, $y_i = 0$ if the policyholder was not at fault or if no claim was made. X is a vector of covariates that is included to control for the risk classification used by the insurer in 2006-2008.

The focus on at fault claims calls for a remark; we only consider at-fault claims where the policyholder was the driver. The reason is that the insurance company does not consider and price by additional drivers. Hence, it is not possible to control for additional drivers in X since these variables aim to explain the policyholders', or equivalently the vehicle owners', risk. If all claims at-culpa are considered there may be a spurious correlation between the error terms resulting from omitted variables regarding the risk classification of additional drivers.[16] For this reason we have sorted out claims at-culpa where the policyholder was not the driver. A sensitivity analysis assesses the implications of this elimination.

C&S argue that the policyholder's probability of owning a certain contract depends on the risk classification X and some random shock ε_i. In a similar way, for any X, the occurrence of an accident at-culpa also depends on some random shock η_i. The error terms are aimed at capturing any residual heterogeneity across agents when the risk classification has been taken into account. The variable of interest is the correlation between the error terms (ρ). If $\rho > 0$ there is an indication of adverse selection since conditional on risk classification, the choice of a contract and the occurrence of an accident are not independent: Contracts with more complete coverage predict a higher probability of an ex post risk.

We extend this by interpreting $\rho < 0$ as an indicator of propitious selection. If policyholders have private information that they are good risks, conditional on risk classification, we expect ex post risk and coverage to be negatively related. One note of caution is the theoretical result in DeDonder and Hindriks (2009), of which we cannot expect a negative correlation. Still, there may exist other measures of good risks that produce a negative correlation in equilibrium. Our null hypothesis is that the error terms ε_i and η_i are not correlated, that is, the choice of a particular contract and ex post risk are independent. Our interpretation of being unable to reject the null is that neither selection dominates, rather than interpreting it as if there is symmetric information.

The second step of the analysis is based on an approach, suggested by Finkelstein and McGarry (2006), to studying the explicit effect of private information about risky traffic behavior for coverage and ex post risk. They argue that this approach provides a more robust test for asymmetric information compared to the correlation test. The reason is that it includes variables that represent the policyholder's private information, which opens up for the possibility of directly observing the effect of private information.

[16] If we consider accidents for driver A where another driver, say B, is at fault we may get a spurious correlation between claims and coverage. The reason is that we condition on the information set related to A and not B. This implies that omitted information about B will affect the correlation coefficient, which may result in a spurious correlation. Getting hit by driver B is a stochastic risk for driver A and this is the reason why A purchases insurance in the first place. Hence the purpose of insurance is to correctly estimate a policyholders' type dependent risk and individuals sharing the same type dependent risk then share the stochastic risk of an accident.

The null of symmetric information is rejected if, conditional on X, private information about traffic behavior is correlated with both insurance coverage and ex post risk occurrence. We test the effect of private information by estimating the following probit models:

$c_i = 1(X\beta_1 + D\beta_2 + \varepsilon_i > 0)$ (3)

$y_i = 1\ (X\delta_1 + D\delta_2 + \eta_i > 0)$ (4)

(i = contract)

The added information compared to equation (1) and (2) is four indicator variables that take the value one if the policyholder has at least one fine for speeding, at least one fine for other traffic offences, one conviction for traffic safety violations, and more than two convictions for traffic safety violations, respectively. The reason why we separate one and several convictions is that we believe that relapsed criminals are higher risks. One conviction may be random, but not several. We also expect that different violations can have different effects due to the social norm.

The coefficients of interest in equation (3) and (4) are β_2 and δ_2. From them we can conclude whether the policyholder's private information about risky traffic behavior has any effect on choosing extensive coverage, and/or the probability of being at fault in a claim. Under adverse selection $\beta_2 > 0$ and $\delta_2 > 0$, which imply that violations of traffic law regulations are associated with more coverage and culpa in claims. A prediction consistent with propitious selection is that higher risk purchases less insurance, implying that the insurers are left with better risks.

D. Controlling for risk classification

Previous studies have pointed out the importance of a careful conditioning on the information set available to the insurance company. The information set is equivalent to all information that is observable and used in premium pricing by the insurance company. However, an important distinction must be made between the information set available to the insurer and the actual risk classification used in premium pricing. The information set is the basis for the actuarial prediction that results in a risk classification. Our preferred approach is therefore to condition on the companies actuarial risk classification. The main reason is that individuals with similar risk classification are considered as homogenous groups by the insurer. A proper implementation of the positive correlation test therefore requires that insurance demand is analyzed across homogenous groups of individuals who likely face the same set of possible insurance contracts. A misspecification may result in a spurious correlation: the accuracy is therefore crucial.

As previously mentioned, our data contains the actuarial prediction of residential area and vehicle risk category, and we control for policyholder age, vehicle age, and driven distance and apply the test to gender and age groups that were previously used in the insurers' actuarial model.[17] The variables in X in all regressions are, age of policyholder, vehicle age, kilometer class, vehicle risk classification and residential area risk classification. We also apply the analysis to the age and gender groups used by the insurer in the actuarial model during 2007 and 2008.

4. Results

A. Replication of previous studies

As discussed earlier, Cohen (2005) did not reject the null of symmetric information for the more experienced driver group since a significant negative correlation was found. We replicate these findings by dividing the policyholders into similar groups. [18] Group one consists of drivers with less than three years of driving experience (N = 15 882): this group has no statistically significant correlation between risk and coverage ($\rho = 0.032$, se = 0.023). The result confirms the results of both Cohen (2005) and C&S.

The second group corresponds to drivers with more than three years of driving experience (N = 340 501): this group has a statistically significant correlation ($\rho = 0.037$, se = 0.009). The

[17] This implies that instead of conditioning on vehicle make, cc, residential area etc., we have access to the actuarial predicted risk regarding vehicle and residential area for the respective insurance types.
[18] Our approach differs in that we focus on culpa-claims and more extensive coverage than Cohen who studies if low-deductible policyholders are associated with more claims. Furthermore, our data does not contain information about driving experience since it is not used in the risk classification. We therefore use a proxy for driving experience by considering age group 18-20 to have less than three years of driving experience and older drivers to have more than three years of driving experience.

results confirm the findings of Cohen in that we reject the null hypothesis, but, in contrast, we find a positive correlation between risk and coverage.[19]

One potential caveat is that the group of inexperienced, or young, drivers is more likely to be homogenous than a sample of different seniority drivers. C&S provide a note of caution when considering individuals with various driving records and ages. One reason is heteroskedasticity, since the distribution of random shocks will depend on seniority, older individual are more likely to have reported a claim due to longer exposure. This potentially biases the correlation test.

B. The standard positive correlation test

Table 2 reports the results from the bivariate probit model of equations (3) and (4) for new policyholders with a vehicle aged 3-20. The total number of observations of new policyholders in 2007 and 2008 with a vehicle in the age interval 3-20 is 295 846.[20]

[19] However, Cohen does not use the lowest deductible in the correlation test. She uses a deductible, referred to as a regular deductible which most of the policyholders in her study choose. The studied insurer offers a low, a regular, a high and a very high deductible. Note that the first provides the most extensive coverage.
[20] In total there are 295 875 observations where the vehicle is in the age interval 3-20. In 29 of the observations the owner is below 18 years: these are excluded from the analysis since they are below the driving license age.

Table 2. Correlation test between All Risk Insurance and Culpa

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Males	Females	Males	Mixed group	Mixed group
Correlation coefficient (ρ)	0.140** (0.061)	0.041 (0.046)	0.015 (0.054)	0.034 (0.040)	-0.007 (0.045)	0.032 (0.039)	0.083*** (0.033)	0.020 (0.031)	-0.004 (0.023)	0.074*** (0.017)
95% Confidence Interval (ρ)	[0.018,0.258]	[-.050,0.131]	[-0.091,0.120]	[-0.045,0.111]	[-0.095,0.081]	[-0.043,0.108]	[0.017,0.148]	[-0.040,0.080]	[-0.050,0.042]	[0.040,0.107]
Log likelihood	-2421.026	-4689.6897	-5982.340	-9009.245	-6960.222	-10101.646	-15387.476	-22013.852	-37969.87	-59579.976
Prob>χ^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Likelihood ratio test of $\rho = 0$, χ^2 (1):	5.083	0.772	0.077	0.703	0.022	0.698	6.155	0.432	0.034	18.748
N	5 669	12 689	10 373	19 026	10 989	18 048	24 686	37 250	63 912	93 204

Notes: The dependent variables are claims at culpa where the policyholder was held either partially or fully responsible, and All Risk Insurance with the low deductible (the highest possible insurance coverage). Independent variables correspond to the insurers' risk classification. The total number of observations of new policyholders in 2006 and 2007 with a vehicle in the age 3-20 is 295 846. Standard errors are in parentheses***, **, * indicate significance at 1%, 5% and 10% levels respectively. The critical value of the chi-square distribution with 1 df is 3.841 at the 95% confidence level. See main text for more details.

Overall it seems that the insurance company is able to handle the information asymmetry problem since there tends to be no significant correlation in the majority of groups. Neither adverse nor propitious selection dominates, except for three groups. Conditional on the risk classification, the correlation coefficient is significant for females in the age group 18-21 at the five percent level, females in the age group 30-39 at the one percent level and for policyholders of both sexes in the age group 50+ at the one percent level.[21]

According to the adverse selection prediction, we expect a significant correlation between higher risk and coverage; hence we expect a significant correlation for young drivers that are generally considered to be higher risks. This is confirmed by the correlation test for females aged 18-21 since we cannot reject the null. On the contrary, the correlation coefficient for young males is insignificant. One explanation is that young individuals refrain from more extensive coverage, since insurance coverage in this age and gender group is often associated with very high premiums. The premium is generally lower for females, so one explanation for the significant correlation for females 18-21 is that they can afford extensive coverage. This may also explain the significant positive correlation in the age group 30-39. Since the premium generally is lower for females, there is more incentive to let the female in the household own and insure the vehicle. Nonetheless, it should be noted that we consider the correlation between coverage and claims where the policyholder was the reported driver: drivers other than the owner are excluded from the analysis. Note, too, that the driver in our data is the stated driver and that the insurer cannot always control for this.

The insignificant correlation for males aged 18-21 calls for a remark: A common problem for Swedish insurers is that young individuals let a parent or another individual with a better rating own the car, the purpose being to reduce the premium. The insurer generally observes a u-shaped relation between more claims and age. The explanation may be that younger individuals use vehicles owned by their parents. If the observed cost turns out to be higher than expected due to a younger driver (a higher risk) being main user, the premium will be too low. Hence, the premium is not actuarially fair.[22] The result regarding the group 50+ may also be a consequence of ownership. Yet again, even though the correlation test is performed

[21] We also apply Finkelstein and Poterba's (2004) approach to test the correlation between coverage and risk; $prob(y=1)=\Phi(X\beta_1+c\beta_2)$ where c= All Risk Insurance with the low deductible. The positive correlation prediction is that $\beta_2 > 0$. This test confirms the results from the positive correlation test.
[22] An actuarial fair premium is a premium that corresponds to the expected cost of the policy.

on at-fault claims where the policyholder is the reported driver, there may be, in some cases, untruthful reports from the driver in the accident.

Sensitivity analysis

We first apply a sensitivity analysis to the vehicle age restriction, since full coverage may not be motivated for older vehicles due to the economic value of the car. We apply the positive correlation test for vehicle age 3-15, 3-10 and 3-5 (see Table 1-3 in the Appendix A). The correlation is insignificant for all groups when the vehicle is 3-5 years old, but becomes significant for females aged 30-39 and the mixed gender group aged 50+ when the vehicle is 3-10 and 3-15 years old. Hence, the correlation structure does not differ a lot when considering different age intervals for the vehicle, and our findings regarding vehicle age 3-20 seem to be robust.

To investigate whether the results are sensitive to the censoring for 2008, we perform a sensitivity analysis of the correlation test on new policyholders for 2007: that is, contracts where we can observe the whole life span. The results indicate that there exists a positive correlation between risk and coverage for females aged 30-39 and the mixed gender age group 50+. The conclusion is that our results regarding new policyholders for 2007 and 2008 do not suffer from a serious under reporting of claims due to the censoring of outcomes of some contracts signed in 2008. See Table 1 in Appendix B.

We also expect the significance level of the correlation coefficient to increase if we consider all claims at culpa. That is, we include cases where a driver other than the owner was at fault in the accident. As previously mentioned, the insurers do not include additional drivers in their risk classification. When including additional drivers, the correlation coefficient also becomes significant for males in the age group 30-39 and the mixed gender group aged 40-49, see Table 4 in Appendix A.

Since the correlation is affected by omitted variables, we expect that the significance level of the correlation coefficients increases if we omit some variables observed by the insurer. To verify this we first apply the correlation test to all reported claims, rather than only at-fault claims, and more extensive coverage. When including all control variables the results suggest a significant positive correlation for all groups, see Table 5 in Appendix A. We do not know, however, whether the results are an effect of claims being reported because policyholders have more extensive coverage, or if they have more extensive coverage because they know that they are likely to report a claim.[23] Still we observe that the correlation coefficient increases as expected if we exclude some of the control variables, see tables 5 and 7 in the Appendix. Second, we test the effect of the correlation test on a mixed gender group aged 21 and older. This group was not used in the actuarial model during 2006-2008.[24] As can be seen in Table 6 in Appendix A, the correlation becomes significant at the one percent level for this group.

One conclusion is that there seems to be a difference in the importance of controls between claims at culpa and claims in general. A change in variables connected to the driver characteristics in the claims at culpa and coverage analysis affect the significance level of the correlation. Similarly, the correlation coefficient increases if variables connected to the vehicles are excluded when testing claims in general and more extensive coverage. Our interpretation is that culpa accidents are more determined by driver characteristics while claims in general are more dependent on vehicle characteristics or random events.[25] Our general conclusion from the sensitivity analysis is that the importance of an accurate conditioning on the insurers' risk classification and the group to whom we apply the test is confirmed.

C. The Finkelstein and McGarry approach

Tables 2 and 3 report the marginal effects from estimating the relationship between private information about risky traffic behavior, more insurance coverage and culpa in equations (3) and (4), respectively. A bivariate probit model is used in groups where there is a significant correlation between equations (3) and (4): similarly, the equations are estimated independently in groups where there is an insignificant correlation.

[23] To investigate if there tends to be an adverse selection effect, a better approach is to compare the outcome in All Risk Insurance with low and high deductibles respectively. An adverse selection prediction is that more claims are reported in the contract with low deductible. To compare the outcome in a setting with low and high deductibles it is necessary to exclude claims that are lower than the highest deductible.
[24] This group consists of the more experienced driver group used in the replication of Cohen (2005).
[25] In unreported regressions we test to exclude vehicle risk classification in the culpa coverage analysis and the correlation structure is not affected. Similarly, there is a small effect of excluding policyholder characteristics in claims in general.

Table 3. Relationship between new policyholders' private information and extensive coverage.

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Males	Females	Males	Mixed group	Mixed group
Speeding	0.253** (0.113)	0.007 (0.007)	0.054*** (0.019)	-0.007 (0.007)	0.061*** (0.020)	0.026*** (0.009)	0.069** (0.032)	0.012* (0.007)	-0.013** (0.006)	-0.025* (0.015)
Other Traffic Offences	-0.198 (0.133)	-0.008 (0.006)	0.011 (0.022)	0.002 (0.007)	-0.012 (0.025)	-0.008 (0.009)	-0.004 (0.043)	-0.016** (0.008)	-0.018*** (0.007)	-0.092** (0.020)
One conviction1973-2007	-	-0.050 (0.019)	-0.119 (0.098)	0.017 (0.019)	0.132 (0.086)	0.004 (0.017)	-0.081 (0.067)	-0.008 (0.009)	-0.023*** (0.007)	0.029* (0.017)
Several convictions1973-2007	0.631*** (0.220)	0.014* (0.009)	-0.033 (0.061)	-0.057*** (0.010)	-0.216*** (0.065)	-0.052*** (0.016)	-0.395*** (0.124)	-0.094*** (0.011)	-0.080*** (0.009)	-0.232*** (0.026)
Correlation coefficient (ρ):	0.143** (0.062)						0.085*** (0.033)			0.077*** (0.017)
95% Confidence Interval (ρ):	[0.021,0.261]						[0.019,0.149]			[0.021,0.261]
Likelihood ratio test of $\rho = 0$, χ^2 (1):	5.240						6.440			20.234
Log likelihood	-2409.5562	-3403.367	-5413.4944	-7769.567	-6248.2716	-9082.3895	-15375.416	-20391.904	35210.911	-59505.089
Prob>χ^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	5 669	12 689	10 373	19 026	10 989	18 048	24 686	37 250	63 912	93 204

Notes: The reported coefficients are marginal effects from probit estimation of (3): a bivariate probit model is used where there is a significant correlation between the residuals. The dependent variable is an indicator of whether the policyholder has All Risk Insurance with low deductible or not. Private information for all groups is represented by four dummy variables taking the value one if the policyholder had one or several on-the spot fines for speeding, one or several on-the-spot-fines for traffic offences, one conviction and two or more convictions for traffic safety violations. The total number of observations is 295 846. ***, **, * represent significance at 1%, 5% and 10% levels respectively. (-) indicates that the variable is omitted due to empty cells[26]. Standard errors are in parentheses. The critical value of the chi-square distribution with 1 df is 3.841 at the 95% confidence level.

[26] The indicator variable for one conviction is always zero for females in age 18-21, which implies that no one in this group had one conviction for traffic safety violations.

Table 4. Relationship between new policyholders' private information and at fault claims.

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Males	Females	Males	Mixed group	Mixed group
Speeding	0.261 (0.169)	0.007* (0.005)	0.003 (0.004)	0.001 (0.002)	0.004 (0.004)	0.000 (0.002)	-0.091 (0.097)	0.002 (0.001)	0.001 (0.001)	0.070* (0.042)
Other Traffic Offences	0.350** (0.156)	0.005 (0.004)	0.011*** (0.006)	0.001 (0.002)	0.001 (0.012)	0.002 (0.002)	0.127 (0.107)	0.003*** (0.001)	0.004*** (0.001)	0.078 (0.053)
One conviction 1973-2007	-	-	-	-0.000 (0.005)	0.012 (0.023)	0.010*** (0.005)	0.299** (0.147)	-0.000 (0.001)	-0.001 (0.001)	0.107*** (0.046)
Several convictions 1973-2007	0.299 (0.333)	0.003 (0.005)	0.001 (0.010)	0.005 (0.004)	0.008 (0.020)	0.005 (0.004)	0.312 (0.237)	0.001 (0.002)	0.001 (0.001)	0.235*** (0.061)
Correlation coefficient (ρ):	0.143*** (0.062)						0.085*** (0.033)			0.077*** (0.044)
95% Confidence Interval (ρ):	[0.021,0.261]						[0.019,0.149]			[0.044,0.110]
Likelihood ratio test of $\rho = 0$, χ^2 (1):	5.240						6.440			20.334
Log likelihood	-2409.5562	-1279.9611	-560.1342	-1225.6187	-700.1885	-1005.2104	-15375.416	-1577.5631	-2702.6105	-59505.089
Prob>χ^2	0.0000	0.0310	0.0122	0.1480	0.6994	0.0349	0.0000	0.0023	0.0003	0.0000
N	5 669	12 689	10 373	19 026	10 989	18048	24 686	37 250	63 912	93 204

Notes: The reported coefficients are marginal effects from probit estimation of (4), a bivariate probit model is used where there is a significant correlation between the residuals. The dependent variable is an indicator of whether the policyholder has reported a claim where s/he is partially or fully at fault. Private information for all groups is represented by four dummy variables taking the value one if the policyholder had one or several on-the spot fines for speeding, one or several on-the-spot-fines for traffic offences, one conviction and two or more convictions for traffic safety violations. The total number of observations of new policyholders for 2006 and 2007 with a vehicle aged 3-20 is 295 846. (-) indicate that the variable was omitted due empty or small cells[27]. Standard errors are in parentheses***, **, * indicate significance at 1%, 5% and 10% levels respectively. The critical value of the chi-square distribution with 1 df is 3.841 at the 95% confidence level.

[27] As in Table 3 the indicator variable for one conviction is always zero for females in the age group 18-21, which implies that no one in this group had one conviction for traffic safety violations. In age groups males 18-21 and females 22-25 the indicator variable for one conviction perfectly predicts cases where the dependent variable is zero. There were 16 and 17 cases respectively where this indicator variable was equal to one, which implies that those with one conviction did not have a reported at fault claim. The indicator for one conviction variable is therefore omitted due to empty cells in these groups.

Table 3 reports the results from estimating the relationship between private information on risky behavior and insurance coverage in equation (3). The results indicate that speeding increases the probability of more insurance, except for the mixed gender and age groups 40-49 and 50+. Moreover, private information about other traffic offences and several convictions for traffic safety violations tend to essentially decrease the probability of more insurance coverage.

Table 4 reports the results from estimating the relationship between private information and at fault claims from equation (4). The results indicate that private information on risky traffic behavior tends to increase the probability of claims where the policyholder was fully or partially at fault. One note of caution is that there may be an under reporting of culpa claims, high-risk drivers who do not purchase extensive insurance have less incentive to report an accident to the insurance company. Whether or not an accident is reported or not is the policyholder’s decision and this is in turn determined by the terms in, and the magnitude of, the contract.

Taken together, the results presented in Table 3 and 4 point to the presence of asymmetric information, also in groups where no significant correlation was found. This implies, in line with the findings of Finkelstein and McGarry (2006), that a test including private information is more revealing in investigating the effect of private information. The results suggest that policyholders with private information are both less and more likely to have extensive insurance, while they have an increased probability of being at fault in a claim.

A potential caveat is that we cannot observe all contracts until they expire since data is censored for 2008. We therefore perform a sensitivity analysis of the effect of private information on culpa where we include only new policyholders 2007, see Table 2-4 in Appendix B. The reason is that the censoring may lead to an under-reporting of culpa claims. The results indicate the same pattern as for new policyholders in 2007 and 2008, the conclusion being that our results are not sensitive to the censoring.

5. Conclusions

A general challenge of any empirical analysis regarding insurance data is the difference in structure across insurance markets. Market heterogeneity may explain why some markets tend to have propitious selection, while others tend to have adverse, or even no selection. It is furthermore reasonable to question whether we should expect to find any evidence of information asymmetries in the insurance market, since an accurate conditioning on the insurer's risk classification is likely to provide an insignificant correlation, at least if the risk classification used by the insurer is efficient. This paper nevertheless shows that being unable to reject the null of zero correlation is not necessarily consistent with symmetric information, or a sufficient risk classification, in the automobile insurance market. When testing the effect of policyholder's private information on traffic safety violations, which is unobservable to the insurer, we find that the market suffers from asymmetric information even in groups where there is no statistically significant coverage-risk correlation.

Our results indicate that policyholders' private information on being a high risk increases the probability of at-fault claims. Furthermore, private information both increases and decreases the probability of extensive insurance coverage. An increase implies an increased risk to the insurer, which increases with the magnitude of insurance coverage since the insurer has to carry a larger share of the economic risk. Similarly, a decrease in the probability of extensive insurance decreases the insurance company's risk. More specifically, the indicator variable for speeding tends to be positively related to extensive coverage, while the indicator for convictions, such as drunken driving and other traffic offences, tends to be negatively related to extensive coverage. The results regarding convictions may further mirror that individuals with prior convictions find themselves less likely to have an accident (Guppy; 1993): with this in mind, it is rational to have a lower demand for insurance. The results regarding speeding is more open for discussion. It could be that the number of speeding tickets is highly correlated with driving experience and distance. In that case it could reflect propitious selection for the older, more experienced groups. This group may perceive themselves as good risks and therefore demand less insurance. However, another possibility is that speeding is correlated with other risk characteristics leading to higher demand for insurance so it could also be interpreted as adverse selection. All in all, our findings suggest that adverse and propitious selection is present simultaneously in the particular market studied. Private information positively related to extensive coverage implies an adverse selection of risks. The reason is that policyholders demand more insurance while they constitute a higher risk. On the contrary, private information negatively related to extensive coverage implies a better selection of risks. The reason is that the high risks select themselves out of the contract.

Our results have important policy implications since they imply that an absence of a risk-correlation is not synonymous with absence of information asymmetries. Policy discussions should consider potential information asymmetries in the specific market, keeping in mind that information that is private in some markets may be public in others. This implies that certain types of information have different implications across markets, countries and even insurance companies. Laws and regulations may also play an important role in whether or not we expect to find evidence of adverse or propitious selection. A reason for considering sub sets of policyholders, rather than a whole population, is that the information between the policyholder and the insurer is not static since the asymmetry likely reduces over time (see Cohen; 2005). Furthermore, there may be different effects of information asymmetries in different subgroups.

Our approach contributes to a potentially viable alternative in testing for information asymmetries. We suggest that future research should consider specific market characteristics and subsets of policyholders that are likely to be affected by information asymmetries. Since adverse or propitious selection can have offsetting effects on the correlation between risk and coverage we find the correlation test plausible if the researcher is interested in ascertaining which selection effect dominates the market. A potential risk with the correlation test, no matter the accuracy of conditioning of the insurers' information set, is that the results are biased by information observed by the insurer and not the researcher.

We suggest that empirical work in this area should not try to find empirical evidence in favor of the adverse selection prediction that generally holds for all markets. While the standard adverse selection assumes one dimension of risk markets tend to be more complex. It might therefore be dangerous to rule out the effect of asymmetric information based on an insignificant risk-coverage correlation. Future research may benefit from interpreting relevant information asymmetries in broader terms than the standard prediction. It is reasonable to believe that ambiguity found across insurance markets does not necessarily imply a contradiction. We rather believe that this is an effect due to market heterogeneity.

References

Aarts L, van Schagen I (2005). ”Driving speed and the risk of road crashes: A review.” *Accident Analysis and Prevention* no 38.

Akerlof, G (1970) "The Market for Lemons: Quality Uncertainty and the Market Mechanism", *Quarterly Journal of Economics*, Vol. 84, No. 3, pp. 488-500.

Bolton, P and M Dewatripont (2005) Contract Theory, Cambridge Mass: MIT Press.

Cawley J and T Philipson (1999) "An Empirical Examination of Information Barriers to Trade Insurance", *American Economic Review*, pp 827-846.

Chiappori P-A. and B. Salanié (1997) "Empirical Contract Theory: The Case of Insurance Data", *European Economic Review* pp 943-950

Chiappori P-A. and B. Salanié (2000) “Testing for Asymmetric Information in Insurance Markets”, *Journal of Political Economy*, vol 108, no1.

Chiappori, P.A. and B. Salanié (2003), "Testing Contract Theory: A Survey of Some Recent Work", in M. Dewatripont, L. Hansen and S. Turnovsky. *Advances in Economics and Econometrics*, Cambridge University Press, Cambridge.

Cohen A (2005), "Asymmetric Information and Learning: Evidence from The Automobile Insurance Market" *The Review of Economics and Statistics*, pp 197-207.

Cohen A (2008), "Asymmetric Learning in Repeated Contracting: An Empirical Study", NBER Working Paper No. 13752.

Cohen and Einav : Cohen A and Einav L (2007) "Estimating Risk Preferences from Deductible Choice", *American Economic Review*, Vol. 97 No. 3, pp 745-788.

Cohen A and P Siegelman (2010), “Testing for adverse selection in insurance markets” *Journal of Risk and Uncertainty* (forthcoming).

Cutler D. and R J. Zeckhauser (2000) "The anatomy of health insurance" Handbook of Health Economics, in: A. J. Culyer and J. P. Newhouse, *Handbook of Health Economics*, edition 1, volume 1, chapter 11.

Cutler D. (2000) “ Health Care and the Public sector” in A Auerbach and M Feldstein (Eds.) *Handbook of Public Economics* (Amsterdam: North Holland 2002).

Dahlby, B. (1983) “Adverse Selection and Statistical Discrimination: An Analysis of Canadian Automobile Insurance” *Journal of Public Economics* 20

Dahlby, B. (1992) “Testing for Asymmetric Information in Canadian Automobile Insurance,” in George Dionne (Ed.) *Contributions to Insurance Economics* (Boston, Kluwer Academic.)

Dionne G., C. Gouriéroux and C. Vanasse (2001) "Testing for Evidence of Adverse Selection in the Automobile Insurance Market: A Comment" *Journal of Political Economy*, University of Chicago Press, vol. 109 No 2.

DeDonder P. and J Hindriks (2009) "Adverse selection, moral hazard and propitious selection" *Journal of Risk and Uncertainty* vol 38, pp 73-86.

DeMeza D., D Webb (2001) "Advantageous Selection in Insurance Markets", *RAND Journal of Economics*, pp 249-262

Fang H, M Keane and D Silverman (2008). Sources of Advantageous Selection: Evidence from the Medigap Insurance Market. *Journal of Political Economy*, University of Chicago Press, vol. 116(2)..

Finkelstein, A and K, McGarry (2006) "Multiple Dimensions of Private Information: Evidence from the Long-Term Car Insurance Market" *American Economic Review*, pp 938-958.

Finkelstein A and J Poterba (2004) "Adverse Selection in Insurance Markets: Policyholder evidence from the U.K Annuity Market" *Journal of Political Economy* Vol 112 no 1.

Forward S. E, Kós-Dienes D and Obrenovic S. (2000) “Invandrare i trafiken, en attitydsundersökning i Värmland och Skaraborgs län” [Immigrants in traffic, a study of attitudes in two counties in Sweden] VTI report 454.

Forward S. E (2006) “The intention to commit driving violations – A qualitative study”. *Transportation Research part F*. 9.

Forward S.E (2008). “Driving violations investigating forms of irrational rationality”. Digital comprehensive summaries of Uppsala dissertations from the faculty of social sciences 44. Uppsala University.

Guppy A (1993), “ Subjective Probability of Accident and Apprehension in Relation to Self-Other Bias, Age, and Reported Behavior”. *Accident Analysis and Prevention*, Vol. 25, No. 4.

Hemenway D (1990) "Propitious Selection", *Quarterly Journal of Economics*, pp 1063-1070.

Jöreskog K.G. Sörbom, D. (1989). ”LISREL 7 - a Guide to the Program and Applications” 2nd Edition. SPSS Publications, Chicago.

Puelz, R. and A. Snow (1994) “Evidence on Adverse Selection: Equilibrium, Signaling and Cross- Subsidization in the Insurance Market” *Journal of Political Economy* 102.

Rotschild M and J E, Stiglitz (1976) "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *The Quarterly Journal of Economics*, MIT Press, vol. 90 No 4.

Salanié B (2005) "The Economics of Contracts: a primer", Cambridge, Mass: MIT Press.

Salanié B (2003) "Testing Contract Theory", CESinfo Economic Studies.

Solomon D. (1964) "Accidents on main rural highways related to speed, driver, and vehicle". Technical report, U.S. Department of Commerce/Bureau of Public Roads.

Åberg L. (1993). “Drinking and driving: intentions, attitudes and social norms of Swedish male drivers.” *Accident Analysis and Prevention* 25.

Åberg L (1998) “Traffic rules and traffic safety”. *Safety Science* 29.

Åberg L and Rimmö P.A. (1998). “Dimension of aberrant driver behaviour”. *Ergonomics* 41.

Appendix A: Sensitivity analysis of vehicle age, claims and omitted control variables

Table 1: Correlation test between All Risk Insurance and Culpa, vehicle age 3-5.

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Females	Females	Males	Mixed group	Mixed group
Correlation coefficient (ρ)	0.307 (0.185)	0.176 (0.198)	-0.023 (0.147)	-0.126 (0.102)	-0.155 (0.107)	-0.037 (0.079)	0.015 (0.068)	0.067 (0.063)	0.005 (0.046)	0.040 (0.034)
Log likelihood	-249.5207	-366.689	-1103.2989	-1661.5886	-1575.6054	-2562.3717	-3933.0821	-6264.1549	-8482.2055	-15204.56
Prob>χ^2	0.0157	0.0033	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	330	560	1 623	2 311	2 280	3 523	6 063	8 960	12 530	22 707

Notes: Dependent variables are claims at culpa where the policyholder was held partially or fully responsible and All Risk Insurance with the low deductible (the highest possible insurance coverage). Independent variables correspond to the insurers' risk classification. Total number of observations are 60 887. Standard errors are in parentheses***, **, * indicate significance at 1%, 5% and 10% levels respectively.

Table 2: Correlation test between All Risk Insurance and Culpa, vehicle age 3-10.

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Males	Females	Males	Mixed group	Mixed group
Correlation coefficient (ρ)	0.092 (0.090)	-0.012 (0.076)	0.064 (0.064)	0.065 (0.051)	-0.019 (0.055)	0.007 (0.006)	0.102*** (0.041)	0.027 (0.038)	0.042 (0.030)	0.093*** (0.022)
Log likelihood	-1266.2341	-1929.264	-4020.8968	-5588.6211	-4880.3022	-6989.9253	-10694.881	-15189.234	-23228.236	-37817.804
Prob>χ^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	1 803	2 925	5 625	8 133	6 700	9 661	15 504	21 408	33 488	55 059

Notes: Dependent variables are claims at culpa where the policyholder was held partially or fully responsible and All Risk Insurance with the low deductible (the highest possible insurance coverage). Independent variables correspond to the insurers' risk classification. Total number of observations are 160 306. Standard errors are in parentheses***, **, * indicate significance at 1%, 5% and 10% levels respectively.

Table 3: Correlation test between All Risk Insurance and Culpa, vehicle age 3-15.

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Males	Females	Males	Mixed group	Mixed group
Correlation coefficient (ρ)	0.078 (0.071)	0.054 (0.050)	0.030 (0.056)	0.045 (0.042)	-0.013 (0.047)	0.027 (0.040)	0.089*** (0.035)	0.028 (0.032)	-0.000 (0.025)	0.077*** (0.018)
Log likelihood	-2031.650	-3680.4919	-5519.0509	-7942.3165	-6509.2132	-9281.796	-14254.525	-20201.478	-33586.787	-51923.579
Prob>χ^2	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.0000	0.0000
N	3 759	7 089	8 384	13 758	9 333	14 435	21 114	30 502	50 174	75 687

Notes: Dependent variables are claims at culpa where the policyholder was held partially or fully responsible and All Risk Insurance with the low deductible (the highest possible insurance coverage). Independent variables correspond to the insurers' risk classification. Total number of observations are 234 235. Standard errors are in parentheses***, **, * indicate significance at 1%, 5% and 10% levels respectively.

Table 4: Correlation test between All Risk Insurance and Culpa claims with all drivers.

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Males	Females	Males	Mixed group	Mixed group
Correlation coefficient (ρ)	0.010* (0.058)	0.052 (0.043)	0.051 (0.047)	0.044 (0.035)	0.018 (0.038)	0.035 (0.033)	0.115*** (0.028)	0.046** (0.024)	0.058*** (0.017)	0.072*** (0.014)
Log likelihood	-2520.1394	-4880.1598	-6186.7039	-9382.2815	-7296.1497	-10567.78	-15954.572	-22954.946	-40651.638	-62459.062
Prob>χ^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	5 669	12 689	10 373	19 026	10 989	18 048	24 686	37 250	63 912	93 204

Notes: Dependent variables are claims at culpa where the policyholder, and drivers other than the owner, was held partially or fully responsible and All Risk Insurance with the low deductible (the highest possible insurance coverage). Independent variables correspond to the insurers' risk classification. Standard errors are in parentheses***, **, * indicate significance at 1%, 5% and 10% levels respectively.

Table 5: Correlation test between All Risk Insurance and all claims.

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Males	Females	Males	Mixed group	Mixed group
Correlation coefficient (ρ)	0.160*** (0.037)	0.158*** (0.027)	0.105*** (0.024)	0.113*** (0.019)	0.099*** (0.022)	0.140*** (0.018)	0.110*** (0.015)	0.133*** (0.012)	0.126*** (0.009)	0.124*** (0.008)
Log likelihood	-3437.3003	-6648.4529	-8418.0474	-12931.521	-9715.8665	-14361.022	-21382.128	-30939.512	-54158.634	-81461.713
Prob>χ^2	0.0000	0.0000	0.0000	0.000	0.000	0.0000	0.0000	0.0000	0.0000	0.0000
N	5 669	12 689	10 373	19 026	10 989	18 048	24 686	37 250	63 912	93 204

Notes: Dependent variables are all reported claims and All Risk Insurance with the low deductible (the highest possible insurance coverage) on insurers' risk classification.

Standard errors are in parentheses***, **, * indicate significance at 1%, 5% and 10% levels respectively.

Table 6: Correlation test between All Risk Insurance and Culpa for a mixed age and group (driver experience > 3 years).

Driver experience > 3	
Correlation coefficient (ρ)	0.037*** (0.009)
Log Likelihood	-222277.62
Prob>χ^2	0.0000
N	340 501

Notes: Dependent variables are claims at culpa where the policyholder was held partially and fully and partially responsible and All Risk Insurance with the low deductible (the highest possible insurance coverage) on insurers' risk classification. The age and mixed gender group is not used by the insurer. Standard errors are in parentheses***, **, * indicate significance at 1%, 5% and 10% levels respectively.

Table 7: Correlation test between All Risk Insurance and all claims with less control variables.

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Males	Females	Males	Mixed group	Mixed group
Correlation coefficient (ρ)	0.214*** (0.034)	0.222** (0.025)	0.177*** (0.022)	0.203** (0.018)	0.149*** (0.020)	0.219*** (0.016)	0.166*** (0.014)	0.202*** (0.012)	0.191*** (0.009)	0.164*** (0.007)
Log likelihood	-3920.1135	-7277.2245	-9543.3883	-14417.56	-10824.715	-16003.139	-24080.2	-34334.118	-60780.65	-89182.44
Prob>χ^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	5 669	12 689	10 373	19 026	10 989	18 048	24 686	37 250	63 912	93 204

Notes: Dependent variables are all reported claims and All Risk Insurance with the low deductible (the highest possible insurance coverage) on insurers' risk classification.

Excluded variables are vehicle risk classification and vehicle age. Standard errors are in parentheses***, **, * indicate significance at 1%, 5% and 10% levels respectively.

Appendix B: Sensitivity analysis of new policyholders 2007 and registered cars.

Table 1: Correlation test between All Risk Insurance and Culpa for new policyholders 2007

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Males	Females	Males	Mixed group	Mixed group
Correlation coefficient (ρ)	0.002 (0.085)	0.025 (0.054)	-0.028 (0.066)	0.035 (0.049)	0.002 (0.056)	0.051 (0.045)	0.110*** (0.040)	0.000 (0.037)	-0.004 (0.028)	0.061*** (0.020)
Log likelihood	-1500.1075	-3095.4631	-3782.6156	-5736.3686	-4405.199	-6570.0699	-9712.517	-14089.007	-24400.295	-37368.932
Prob>χ^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	3 489	7 753	6 420	11 736	6 933	11 273	15 542	23 379	40 200	57 987

Notes: Dependent variables: claims at culpa where the policyholder was held partially or fully responsible and All Risk Insurance with the low deductible (the highest possible insurance coverage). Independent variables correspond to the insurers' risk classification. Standard errors are in parentheses***, **, * indicate significance at 1%, 5% and 10% levels respectively.

Table 2. Relationship between new policyholders' private information and culpa.

Age:	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Mixed group	Females	Males	Mixed group	Mixed group
Speeding	0.019* (0.014)	0.005 (0.006)	0.007 (0.005)	0.001 (0.003)	0.009* (0.006)	0.002 (0.003)	-0.139 (0.120)	0.002 (0.002)	0.001 (0.001)	0.073 (0.050)
Other Traffic Offences	0.025** (0.017)	0.008* (0.005)	0.006 (0.006)	0.003 (0.003)	0.009 (0.008)	0.005* (0.003)	0.162 (0.124)	0.005*** (0.002)	0.005*** (0.002)	0.112* (0.050)
One conviction 1973-2007	-	-	-	0.003 (0.007)	-	0.006 (0.006)	0.278* (0.172)	-0.000 (0.001)	-0.001 (0.002)	0.142*** (0.053)
Several convictions 1973-2007	0.038 (0.040)	0.002 (0.007)	0.008 (0.019)	0.007 (0.006)	-	0.012** (0.007)	0.522** (0.260)	-0.001 (0.002)	0.001 (0.001)	0.294*** (0.071)
Correlation coefficient (ρ)							0.114*** (0.040)			0.065*** (0.021)
Log likelihood	-345.27846	-931.24069	-387.46899	-4884.304	-444.22095	-745.42265	-9698.8307	-1110.0727	-1945.9379	-37317.798
Prob>χ^2	0.0102	0.0186	0.00215	0.1004	0.4312	0.0352	0.0000	0.0015	0.0000	0.0000
N	3 489	7 753	6 420	11 736	6 993	11 273	15 542	23 379	40 200	57 987

Notes: The reported coefficients are marginal effects from a probit estimation of equation (4) for new policyholders in 2007. A bivariate probit estimation of equations (3) and (4) is used where there is a significant correlation between the residuals. Private information for all groups is represented by four dummy variables taking the value one if the policyholder had one or several on-the spot fines for speeding, one or several on-the-spot-fines for traffic offences, one conviction and two or more convictions for traffic safety violations.***, **, * represents significance at 1%, 5%, 10% levels respectively. (-) indicate that the variable was omitted due to empty or small cells. Standard errors are in parentheses.

Table 3. Relationship between new policyholders' private information and coverage.

Age	18-21		22-25		26-29		30-39		40-49	50+
Gender	Females	Males	Females	Males	Females	Males	Females	Males	Mixed group	Mixed group
Speeding	0.060*** (0.030)	0.005 (0.009)	0.103*** (0.025)	-0.014 (0.009)	0.086*** (0.025)	0.031*** (0.004)	0.036 (0.040)	0.008* (0.009)	-0.034*** (0.009)	-0.026 (0.019)
Other Traffic Offences	-0.043** (0.018)	-0.010 (0.007)	0.026 (0.028)	0.001 (0.009)	-0.024 (0.031)	-0.022 (0.012)	0.024 (0.053)	-0.013 (0.010)	-0.016** (0.008)	-0.080*** (0.025)
One conviction 1973-2007	-	-0.054 (0.019)	-0.204* (0.090)	0.014 (0.022)	0.345*** (0.098)	0.019 (0.023)	-0.120*** (0.082)	-0.012 (0.011)	-0.028*** (0.009)	-0.030 (0.021)
Several convictions 1973-2007	0.273*** (0.099)	0.035*** (0.013)	0.056 (0.086)	-0.065*** (0.014)	-0.206* (0.097)	-0.035*** (0.022)	-0.567*** (0.172)	-0.115*** (0.014)	-0.080*** (0.012)	-0.229*** (0.033)
Correlation coefficient (ρ)							0.114*** (0.041)			0.065*** (0.021)
Log likelihood	1139.8335	-2155.7906	-3381.2977	-4884.304	-3945.5089	-5813.1039	-9698.8307	-12940.977	-22408.904	-37317.798
Prob>χ^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	3 489	7 753	6 420	11 736	6 933	11 273	15 542	23 379	40 200	57 987

Notes: The reported coefficients are marginal effects from a probit estimation of equation (3) for new policyholder in 2007 only. A bivariate probit estimation of (3) and (4) is used where there is a significant correlation between the residuals. Private information for all groups is represented by four dummy variables taking the value one if the policyholder had one or several on-the spot fines for speeding and one or several on-the-spot-fines for traffic offences, one conviction and two or more convictions for traffic safety violations.***, **, * represents significance at 1%, 5%, 10% levels respectively. (-) indicate that the variable was omitted due to empty cells. Standard errors are in parentheses.