Asymmetric Information in Automobile Insurance: Evidence from Driving Behavior[*]

Alois Geyer[†] Daniela Kremslehner[‡] Alexander Muermann[§]

Based on a unique data set of driving behavior we test whether private information in driving characteristics has significant effects on contract choice and risk in automobile insurance. We define a driving-factor based on overall distance driven, number of car rides, and speeding. Using local weather conditions, we account for the endogeneity of the driving-factor. While this driving-factor has an effect on risk, there is no significant evidence for selection effects in the level of third-party liability and first-party insurance coverage.

Keywords: private information; automobile insurance; driving behavior; telematic data; pay-as-you-drive insurance; endogeneity; sample selection

[*] We are grateful to three anonymous reviewers, Jeffrey Brown, Keith Crocker, Randy Dumm, Martin Eling, Ruediger Fahlenbrach, Robert Hauswald, Robert Kremslehner, Christian Laux, Christian Leuz, Henri Loubergé, Raphael Markellos, Olivia Mitchell, Charlotte Østergaard, Stefanie Peer, Daniel Sturm, and seminar participants at BI Norwegian Business School, Temple University, University of Georgia, Wharton School, EPFL Lausanne, the EGRIE meeting, the AEA annual meeting, the NBER Insurance Workshop, the Risk Theory Society meeting, the ARIA meeting, the European Winter Finance Summit, and the Annual Congress of the German Insurance Science Association for their helpful comments and suggestions.

[†] Department of Finance, Accounting and Statistics, Vienna University of Economics and Business, and VGSF (Vienna Graduate School of Finance), Welthandelsplatz 1, Building D4, A-1020 Vienna, Austria, alois.geyer@wu.ac.at

[‡] Department of Finance, Accounting and Statistics, Vienna University of Economics and Business, Welthandelsplatz 1, Building D4, A-1020 Vienna, Austria, daniela.kremslehner@wu.ac.at

[§] Department of Finance, Accounting and Statistics, Vienna University of Economics and Business, and VGSF (Vienna Graduate School of Finance), Welthandelsplatz 1, Building D4, A-1020 Vienna, Austria, alexander.muermann@wu.ac.at

Electronic copy available at: https://ssrn.com/abstract=2850537

1 Introduction

This paper provides new insights into the relevance of private information in insurance markets based on a telematic data set of insured cars that is inaccessible to the insurance company.[1] The data set contains detailed information about driving behavior[2] (e.g., speed, distance driven, and road type) for policyholders who opted for a pay-as-you-drive contract. While the insurance company uses the aggregate distance driven for the premium calculation, it contractually refrains from accessing any other telematic data.[3] In addition, we also have access to the corresponding insurance data set which includes all variables used for pricing, policyholders' contract choice (third-party liability and first-party coverage), and information about the submission of liability claims. We link this insurance data set to the telematic data set on the car level. As all information, except distance, contained in the telematic data is unobserved by the insurance company, we can directly test whether private information about driving behavior is relevant for and how it is linked to the policyholder's choice of the insurance contract and the conditional loss distribution.

Controlling for the risk classification of the insurance company, we test whether driving characteristics have an effect on the choice of third-party liability and first-party coverage and/or an effect on a subsequent downgrade of the Bonus-Malus class.[4] We use the overall distance driven, the number of car rides, the interaction between the two, and a speeding index to define a driving-factor. These driving characteristics may both determine contract choice (selection effect) and be affected by contract choice (incentive effect), and thus be

[1] *Telematics* stands for the fusion of *telecommunication* and *informatics*. It is typically based on a GPS device which allows for the transmission of information about moving objects, e.g., as used in navigation systems.

[2] Data is recorded approximately every two kilometers (1.24 miles) by a telematic device which is installed in the insured car.

[3] The production and installation of the hardware into the cars as well as the collection and management of the telematic data is carried out by an independent telematic company.

[4] Premiums for third-party liability insurance are based on an experience rating system. A downgrade of the Bonus-Malus class is triggered by the submission of at least one liability claim during one year and results in a higher premium for the following year. We use such a downgrade in the year following the beginning of our telematic data as a proxy for risk.

Electronic copy available at: https://ssrn.com/abstract=2850537

endogenous. We use weather conditions (including precipitation, temperature, snowfall) observed after contract choice but while driving is recorded as instruments. Using detailed telematics data we identify the main location of the car and match it locally with information from 150 weather stations. Thereby we make it possible to disentangle selection effects and incentive effects.

Our results show that both overall distance (as observed by the insurance company) and the driving-factor mentioned above (based on additional private information that is not accessible to the insurance company) increase the likelihood of a downgrade of the Bonus-Malus class. The coefficient's test statistic is close but below the usual critical values, though.[5] These driving characteristics do not affect policyholders' contract choice.[6] Thus, while distance and the driving-factor are plausible and potential risk factors, we find insufficient evidence for relevant selection effects with respect to insurance coverage.

Most of the empirical literature on asymmetric information in insurance markets analyzes insurance data alone and tests for the sign of the correlation between the level of insurance coverage and ex-post realizations of risk controlling for the risk classification of the insurance company. The classical models both of adverse selection and moral hazard (Arrow, 1963; Pauly, 1974; Rothschild and Stiglitz, 1976; Harris and Raviv, 1978; Holmstrom, 1979; Shavell, 1979) are based on one-dimensional private information and predict a positive correlation. This prediction has been confirmed in the health insurance market (Cutler and Reber, 1998; Cutler and Zeckhauser, 1998) and in the annuity market (Finkelstein and Poterba, 2004, 2014; McCarthy and Mitchell, 2010). However, there is also evidence for a negative correlation between the level of insurance coverage and claims probability in the markets for life insurance (Cawley and Philipson, 1999; McCarthy and Mitchell, 2010) and

[5] Robustness checks below show that excluding speeding from the set of driving characteristics increases both the coefficient and its statistical significance on a downgrade of the Bonus-Malus class. This suggests that distance and the number of car rides are more important risk factors than speeding.

[6] The signs show an insignificant negative effect on the level of third-party liability coverage, indicating advantageous selection, and an insignificant positive effect on the level of first-party coverage, indicating adverse selection.

Electronic copy available at: https://ssrn.com/abstract=2850537

for Medigap insurance (Fang et al., 2008). Moreover, no statistically significant correlation has been found in automobile insurance (Chiappori and Salanié, 2000; Dionne et al., 2001; Cohen, 2005)[7] and in long-term care insurance (Finkelstein and McGarry, 2006).[8]

Chiappori et al. (2006) examine the extent to which models of adverse selection and moral hazard can be generalized while still predicting a positive correlation between the chosen level of insurance coverage and the expected value of indemnity.[9] They emphasize that hidden degree of risk aversion can be pivotal for violating the prediction of positive correlation. de Meza and Webb (2001) show that a separating equilibrium with a negative relation between coverage and accident probability can exist if hidden information about the degree of risk aversion is combined with hidden investment in risk reduction, and if insurance contracts entail administrative costs. Finkelstein and Poterba (2014) also argue that if asymmetric information is present on multiple characteristics, including the degree of risk aversion, then the result of rejecting (not rejecting) the hypothesis of non-dependence between the level of insurance coverage and risk may not be indicative of the existence (absence) of asymmetric information. Cohen and Einav (2007) develop a structural model which accounts for unobserved heterogeneity in both risk and risk aversion. By using a large data set of an Israeli insurance company, they find that unobserved heterogeneity in risk aversion is much larger than unobserved heterogeneity in risk. Sandroni and Squintani (2013) study an equilibrium model with overconfident policyholders and find that unobservable overconfidence can explain the negative relationship between the level of insurance coverage and ex-post realizations in a competitive market. We refer to Cohen and Siegelmann (2010) for a review of the empirical literature on asymmetric information in insurance markets.

[7] These papers only examine first-party coverage while we test for selection effects in third-party liability and first-party coverage.

[8] Puelz and Snow (1994) did find a positive relation between coverage and risk. Their result, however, was subsequently challenged by Chiappori and Salanié, 2000, and Dionne et al., 2001. While Cohen (2005) did not find any correlation for beginning drivers, she did find a statistically significant positive relation for experienced drivers.

[9] If there are multiple loss levels, Koufopoulos (2007) shows that the positive correlation property between the level of insurance coverage and *accident probability* (as opposed to the expected value of indemnity) may not hold.

Electronic copy available at: https://ssrn.com/abstract=2850537

We test for the correlation of the generalized residuals based only on the insurance data alone and fail to reject the null hypothesis of zero residual correlation between the level of insurance coverage (both third-party liability and first-party) and a subsequent downgrade of the Bonus-Malus class. By adding distance and the driving-factor to the model, we still fail to reject the null hypothesis of zero residual correlation, consistent with the direct evidence that there are no selection effects based on driving characteristics.

Our paper is most closely related to the recent literature that tests for the effects of multidimensional private information in insurance markets. Finkelstein and McGarry (2006) use individual-level survey data on long-term care insurance and show that individuals' self-reported beliefs of entering a nursing home are positively related to both subsequent nursing home use and insurance coverage. Despite the existence of this risk-based selection, actual nursing home use and insurance coverage are not positively correlated. The authors explain this fact by providing evidence that the risk-based selection is offset by a selection based on heterogeneous degrees of risk aversion as proxied by seat belt usage and investment in preventive health care measures. Fang et al. (2008) also use individual-level survey data on Medigap insurance to examine the reasons for the significant negative correlation between insurance coverage and medical expenditure. They show that cognitive ability rather than risk preferences is the essential factor explaining this negative relation. Robinson et al. (2018) use survey data of drinkers and drivers including risk preferences, driving habits, and subjective assessments of accident risk. They provide evidence that risk-tolerant policyholders demand less insurance coverage. This negative correlation between risk-tolerance and coverage is offset by moral hazard, as the overall correlation between ex-post accident risk and insurance coverage is zero.

We contribute to this literature by analyzing a unique data set that is provided by an independent and unbiased third party, the telematic company. The data contains detailed information about real decisions and behavior of individuals that is of direct interest to but

Electronic copy available at: https://ssrn.com/abstract=2850537

unobserved by the insurance company.[10] The telematics data allows us to derive several aspects of driving behavior, and test for their relations to contract choice and risk.

Finkelstein and Poterba (2014) propose an empirical test based on "unused observables," i.e., on characteristics which are observed by the insurance company but are not used for pricing, either voluntarily or for legal reasons. They argue that if those characteristics are significantly related to contract choice and risk, then this is direct evidence of relevant private information which is not confounded by hidden information on risk preferences. In their study of the UK annuity market, they use postcode information which is collected by the insurance company but not used for pricing. They find that the inhabitants' socio-economic characteristics of different postcode areas are correlated with both survival probability and choice of insurance coverage. Similarly, Saito (2006) uses postcode information which is collected but not used by insurance companies for pricing in automobile insurance. The author rejects the hypothesis that policyholders who live in high accident probability regions are more likely to purchase insurance. Unused but observed data, although not used in pricing, might be used in other types of underwriting activities by the insurance company. For example, policyholders who observably differ in their underlying risk might be offered different contracts, might be scrutinized differently in the claims settlement process, or might face different renewal or cancellation policies. In that case, a significant relation between the "unused observable" and contract choice might reflect those different underwriting policies. The telematic data set provides us with information which is *unobserved* by the insurance company. Thus, the insurance company is not able to condition any type of underwriting or cancellation activity on that information.

Most of the empirical literature testing for the effects of private information in insurance markets is based on a data set of a single insurance company. Salanié (2017) highlights that one has to be cautious in interpreting the results within a market equilibrium. For example,

[10] Responses to survey questions can be biased, in particular, if they relate to self-reported probabilities of future events. Examples include the anchoring bias of unfolding bracket questions (Hurd et al., 1998; Hurd, 1999) and problems of focal responses (Gan et al., 2005).

Electronic copy available at: https://ssrn.com/abstract=2850537

if different insurers specialize in attracting different risk types with different contracts, then we might not observe selection effects within a single insurer, but there might be selection effects across different insurers. While our analysis is also based on data of a single insurer, this insurer was the first and only company offering a pay-as-you-drive contract to the overall market at that time. Thus, our data is the market-wide data of pay-as-you-drive insurance contracts.

Last, our setting further benefits from the fact that liability insurance is mandatory and policyholders who are rejected by insurers are distributed evenly among all insurance companies in the market. This is particularly important as Hendren (2013) finds more private information held by individuals who are rejected by insurance companies compared to nonrejectees. This can explain the lack of significant results of previous literature on the existence of private information in insurance markets.

The paper is structured as follows. Section 2 provides detailed information about the pay-as-you-drive insurance contract, and about the telematic and insurance data sets. In section 3 we specify the econometric model and provide details on how we account for endogeneity and sample selection using weather conditions as instruments. We present empirical results, perform robustness tests, and discuss our main findings in section 4.

2 Background and Data

The insurance company offers a pay-as-you-drive insurance contract in addition to its existing car insurance contract. Cars insured under this contract are equipped with a telematic device which uses GPS. The pricing of this pay-as-you-drive contract is based on the aggregate distance driven – fewer kilometers driven imply a lower premium – and on the road type used.[11] The company distinguishes between three road types: urban, country road, and

[11] For a total distance of up to 4,000 km (2,485 miles) per year the premium for liability and comprehensive insurance is reduced by 25%, between 4,000 km (2,485 miles) and 6,000 km (3,728 miles) by 20%,

Electronic copy available at: https://ssrn.com/abstract=2850537

motorway. The distance driven on country roads and motorways is scaled down by a factor of 0.8. Furthermore, policyholders who choose the pay-as-you-drive contract get a 5% discount on the premium of full comprehensive insurance coverage. In addition to the pay-as-you-drive feature, the telematic device is equipped with an emergency device and a crash sensor. If activated, either by the car driver or in case of an accident, an emergency signal is sent to the helpdesk of the insurance company. The helpdesk will then try to contact the policyholder and call emergency services if needed or if the policyholder cannot be reached. An additional benefit of the telematic device is that stolen cars can be tracked via GPS. Policyholders pay a one-time fee for the installation of the telematic device and a monthly fee for the safety services.

Since policyholders can choose this pay-as-you-drive contract, analyzing the relations between coverage choice, risk, and driving behavior (as well other factors) using only data for those who opted for this contract may lead to a sample-selection bias. We use Heckman's approach[12] to account for a potential selection bias, using additional data for policyholders who did not choose this contract. The empirical results show that the pay-as-you-drive contract is more likely to be chosen by younger, female policyholders living in urban and/or wealthier areas, who drive old(er) and/or more valuable cars with less engine power.

The economic rationale for pay-as-you-drive insurance contracts is the internalization of accident and congestion externalities. Edlin and Karaca-Mandic (2006) estimate that the externality cost due to an additional driver in California is around $1,725 to $3,239 per year. Vickrey (1968) proposed the idea of *distance-based pricing* as a solution to the externality problem. In the U.S., many insurance companies, e.g. Progressive, Allstate, and State Farm, offer pay-as-you-drive insurance contracts for privately owned cars. Liberty Mutual offers pay-how-you-drive insurance contracts for fleets. Edlin (2003) argues that monitoring costs

between 6,000 km (3,728 miles) and 8,000 km (4,971 miles) by 15% and between 8,000 km (4,971 miles) and 10,000 km (6,214 miles) by 10%.

[12] We use a modified version of Heckman's approach because we also address endogeneity in the equations of interest.

Electronic copy available at: https://ssrn.com/abstract=2850537

for mileage-based pricing might be too high and suggests that regulatory enforcement could be necessary since private gains might be much smaller than social gains. Bordoff and Noel (2008) estimate that a US nationwide implementation of pay-as-you-drive insurance would result in a 8% reduction of mileage driven which would yield a social benefit of $50 billion per year, a reduction of carbon dioxide emission by 2%, and a reduction of oil consumption by 4%. They also estimate that two thirds of all households would pay a lower premium under pay-as-you-drive insurance with an average saving of $270 per car per year. Parry (2005) shows that the welfare gains of implementing pay-as-you-drive insurance in reducing driving-related externalities are much larger compared to the welfare gains obtained by increasing gasoline tax.

Due to the safety features of telematic devices, the European Commission has passed a recommendation supporting the EU-wide implementation of a telematic based emergency call (eCall) service for the transmission of in-vehicle emergency calls (European Commission, 2011). This service is mandatory for all new cars since April 2018. Several automobile manufacturers equip their cars with telematic units and offer various additional services to their customers, e.g., automatic crash response and stolen vehicle tracking.

2.1 Telematic Data

An independent telematic company develops the hardware and collects and manages the telematic data. Each data point includes date, time, GPS-coordinates, direction of driving, current speed, distance driven since the last data point, ignition status of the engine, and road type (urban, country road, or motorway). A data point is recorded when the engine is started, after approximately every two kilometers (1.24 miles) driven, and when the engine is switched off. Our data set covers 2,340 cars for a period of 3 months, from February 1st, 2009, to April 30th, 2009, comprising 3.7 million individual data points.

We restrict the data set to car rides where a definitive start and end was recorded. Moreover,

Electronic copy available at: https://ssrn.com/abstract=2850537

we exclude car rides with unrealistically high values of speed (above 200 km/h = 124.27 mph which is above the 99.9% quantile of the empirical distribution) and of distances between data points (above the 99.9% quantile). The excluded car rides are likely to result from a connection failure with the GPS satellite.[13] These exclusions leave us with 3.15 million data points. Table 1 displays the summary statistics of the telematic data. The average total distance driven by each policyholder within the three months period is 2,061 km (1,281 miles), which translates into a yearly average of 8,212 km (5,103 miles). Based on data from the local automobile club, the nationwide average yearly distance driven per driver is 13,140 km (8,165 miles). Thus, policyholders of the telematic insurance contract drive less than the nationwide average. This suggests a selection effect which may be due to the distance-dependent discount offered by the contract or other effects. As noted above, we are going to account for a potential selection bias.

In order to measure the risk type of an insured person we use the total distance driven, the number of car rides, the interaction among the latter two,[14] and average speeding above legal speed limits. This speeding index is given by

$$\text{speeding} = \frac{\sum_j \sum_{i \in \Delta_n} (v_{ij} - u_j)}{n}, \tag{1}$$

where j is the road type (urban, country, motorway), u_j is the countrywide legal speed limit for road type j in km/h (urban: 50 km/h = 31.07 mph, country: 100 km/h = 62.14 mph, motorways: 130 km/h = 80.78 mph), $i = 1, ..., n$ is a data point, v_{ij} is the speed of the car at data point i on road type j, and $\Delta_n = \{i = 1, \dots, n | v_{ij} > u_j\}$ is the set of data points at which the speed of the car is above the legal speed limit.[15]

[13] Most of those excluded car rides reveal further unrealistic characteristics such as speed above 200 km/h (124 mph) at the time the engine is switched on or off, or at the only data point in between, or in urban areas.

[14] This interaction accounts for various driving patterns (e.g. an urban driver typically drives more frequently but shorter distances) and experience (e.g. a policyholder who drives a lot of long-distance car rides may have more driving experience.)

[15] The countrywide legal speed limits are the maximum speed limits. The actual legal speed limit can be lower either permanently, e.g., in residential areas or road sections prone to accidents, or temporarily,

Electronic copy available at: https://ssrn.com/abstract=2850537

Table 1: Summary statistics of telematic data

		road type			
		urban	country	motorway	all
number of cars					2,340
number of car rides					537,181
speed	mean	29.7	45.9	70.4	48.5
(mph)	stdev	11.8	14.0	14.9	22.2
distance/ride	mean				5.6
(miles)	stdev				11.3
distance/car	mean	544	325	496	1,281
(miles)	stdev	461	394	686	1,221

The insurance company has access only to the telematic data that is necessary for pricing the pay-as-you-drive contract, i.e., to the aggregate distance driven per road type. The insurer contractually refrains from accessing any other telematic data because of privacy concerns. The telematic data set thus provides us with detailed private information about driving behavior which is inaccessible to the insurance company. This setting allows us to directly test whether private information as reflected in driving behavior is relevant for the level of insurance coverage and risk.

2.2 Insurance Data

For all privately owned cars in the telematic data set the corresponding insurance contract data is linked on the car level via an anonymous identification number. Corporate cars are excluded from the data set because insurance coverage is not chosen by the driver but by the corporation, and driving behavior may not be comparable to that of private cars. The insurance data comprises all the information used for pricing of the policies at the beginning of February 2009. An update of the insurance data set for February 2010 is used to extract information about the submission of a liability claim during that year. We therefore restrict the telematic data set to those cars which are still insured under the pay-as-you-drive

e.g., due to road works or construction sites. We thus might underestimate the extent to which drivers speed by using the countrywide legal speed limits for each road type.

Electronic copy available at: https://ssrn.com/abstract=2850537

contract after one year. Last, only cars with more than 4 kW (5.4 HP) were included[16], and motorhomes have been excluded. This leaves us with 1847 insurance contracts for analyzing the role of private information on contract choice and risk.

For each contract we observe or derive the following information to be used in the empirical analysis[17]:

- *Car-related information:* age, engine power, and catalog price of the car at initial registration.

  Since the catalog price may become less informative with increasing age of the car, we convert the catalog price into an 'adjusted price' by using the discount factor $\exp\{-0.2{\cdot}\text{age}\}$.[18]

- *Policyholder-related information:* age, gender, postcode, and purchasing power.

  The postcode information is used to identify whether the policyholder lives in a city which can be characterized as urban or not. However, the policyholder is not necessarily the primary driver of the car.[19] In order to determine whether the driver of the car lives in an urban area, we use the GPS-coordinates to first identify the most frequent parking position of a car. We label this position as the 'main location,' indicating the location where most of the trips start and/or end. In a second step, we identify the (nearest) zip-code of the main location and use a threshold of 40,000 inhabitants (according to the zip-code) to define a dummy-variable 'urban.'

  The insurance company does not collect income information in the underwriting process. To control for income in our analysis, we use aggregate data on purchasing power

[16] All cars with 4 kW (5.4 HP) or less are micro-cars which are license-exempt vehicles with a maximum speed of 45 km/h (28 mph). The driving behavior of a micro-car is closer to the driving behavior of a moped than to that of a car.

[17] This information is available for the 1847 clients who opted for the pay-as-you-drive contract as well as for 1987 clients who did not (except for location-based data).

[18] This resolves the issue that the catalog price is highly correlated (above 0.9) with engine power, while the correlation of the adjusted price with the catalog price is about 0.7, and with the car's age about -0.65.

[19] The insurance contract covers every person that drives the car with the approval of the policyholder.

Electronic copy available at: https://ssrn.com/abstract=2850537

for 2009.[20] It is defined as yearly gross income minus direct taxes and social security contributions plus interest earnings and transfer payments in terms of an average on the postcode level. We merge the average purchasing power per resident with the insurance data set through the postcode information.

- *Bonus-Malus rating information:* Premiums for third-party liability insurance are based on an experience-rating scheme.[21] There are 19 Bonus-Malus classes which reflect the car owner's history of claims. Each Bonus-Malus class is related to a scaling factor of a base premium ranging from 44% (lowest class) to 170% (highest class). A car owner with no driving experience starts with 110% of the base premium. If a policyholder does not file a liability claim during a year, then she is upgraded one class (*Bonus*) and pays the next lowest percentage of the base premium in the following year. If a policyholder files a liability claim during the year, then she is downgraded three classes (*Malus*) and pays the corresponding higher percentage in the following year.[22] We use downgrades of the Bonus-Malus record to proxy for ex-post risk (see below).

  Considering the current Bonus-Malus class alone does not account for the rating impact and the associated effects on the premium, which differs depending on the rating class prior to the down- or upgrade. The (positive or negative) effects on the premium are a non-monotonic function of the current level. We account for this fact in terms of two different variables, measuring the impact on the premium in percentage points (positive for downgrades, and negative for upgrades).

- *Coverage choice:*

  The insurance company offers three levels of first-party coverage: none, comprehensive insurance (covers losses from vandalism, theft, weather etc.), and full comprehensive

[20] The purchasing power data was provided by the Austrian Institute for SME Research.

[21] Traffic violations do not affect the rating.

[22] The national insurance association monitors the Bonus-Malus record for each nationwide registered car owner which is accessible to all insurance companies.

Electronic copy available at: https://ssrn.com/abstract=2850537

insurance (in addition including at-fault collision losses).[23] In the empirical analysis we distinguish contracts which cover at-fault losses (full comprehensive insurance) from those that do not.

Two levels of third-party liability coverage are offered by the insurance company, which are both in excess of the minimum level of € 6 million mandated by the insurance law: € 10 million ($13.9m) and € 15 million ($20.9m).[24] In the empirical analysis we distinguish contracts with a high limit from those with a low limit.

Table 2 provides the summary statistics of car- and policyholder-related data. Full comprehensive and high liability coverage is on average bought by older policyholders and for more recently built and more valuable cars with a stronger engine. Moreover, customers of full comprehensive and high liability coverage have a better Bonus-Malues rating.

3 Econometric Model

3.1 Econometric approach

We test for the direct effect of private information on contract choice and risk by extending the econometric model suggested by Finkelstein and Poterba (2014). Their model is based on Chiappori and Salanié (2000) who propose the following probit model for insurance coverage and risk

$$\text{Coverage} = 1(X\beta_c + \varepsilon_c > 0) \tag{2}$$

$$\text{Risk} = 1(X\beta_r + \varepsilon_r > 0), \tag{3}$$

[23] We do not use information on deductible choice since the standard deductible of € 300 ($418) is chosen by more than 99% of all policyholders.

[24] These levels of third-party liability coverage are representative for Europe but very high compared to those offered in the U.S. Policyholders in the U.S. may purchase additional liability coverage through personal umbrella policies.

Electronic copy available at: https://ssrn.com/abstract=2850537

Table 2: Summary statistics of car- and policyholder-related data

		pay-as-you-drive		3rd party coverage		1st party coverage		BM rating downgrade	
		no	yes	low	high	other	full	no	yes
speeding	mean		3.1	3.2	3	3.1	3.2	3.1	3.2
	stdev		2.1	2.1	2	2	2.1	2	2.3
drives/day	mean		3.6	3.7	3.6	3.5	3.7	3.6	3.5
	stdev		2.2	2.2	2.1	2.2	2.2	2.2	2.1
distance/day (miles)	mean		19.4	19.4	19.4	17.6	20.3	19.6	17.3
	stdev		14.5	14.9	13.5	14.3	14.5	14.6	12.8
engine (HP)	mean	60.6	65.0	64.6	65.8	62.3	66.2	65.0	63.8
	stdev	24.4	27.8	28.0	27.1	24.3	29.2	28.1	23.5
value (1000$)	mean	19.5	22.3	21.7	23.7	12.8	26.7	22.7	17.4
	stdev	15.7	16.8	17.5	15.2	13.4	16.6	17.1	13.4
age car (years)	mean	4.4	3.5	3.7	3	6.7	1.9	3.3	4.9
	stdev	5	3.7	3.9	3.2	4.2	2.1	3.6	4
age (years)	mean	51.8	48.7	48.1	49.9	48.1	48.9	48.6	49.2
	stdev	14.7	15.1	14.9	15.5	15.7	14.8	15.1	15
purchasing power (1000$)	mean	23.8	24.5	24.5	24.4	24.2	24.6	24.5	24.4
	stdev	3.6	3.9	4.0	3.6	3.8	3.9	3.9	2.9
male (%)	mean	64.4	61.2	60.9	61.8	61.2	61.2	61	62.9
urban (%)	mean	32	42.1	43.1	39.8	40	43.1	41.9	45
BM rating (%)	mean	48.7	51.9	52.2	51.2	54.7	50.6	51.8	53
impact downgrade (pp)	mean	15.6	17	17.3	16.3	18	16.5	16.9	17.6
impact upgrade (pp)	mean	-1.3	-1.9	-2.1	-1.7	-2.5	-1.7	-1.9	-2.1
BM rating downgrade (%)	mean		7.6	7.8	7	9.6	6.6	0	100

Notes: male, urban and BM rating downgrade are binary variables; their means are expressed in percentage terms; the impact of down- and upgrades is given in percentage points (pp); statistics are based on 1847 observations for policyholders who opted for the pay-as-you-drive contract, except for the first column which is based on 1987 clients who did not.

where X is the vector of all risk classifying variables used by the insurance company. Chiappori and Salanié (2000) interpret a non-zero correlation ρ between the error terms ε_c and ε_r as an indication for the existence and the effect of private information. A statistically significant positive correlation coefficient is consistent with the classical models of adverse selection and moral hazard with asymmetric information about one parameter of the loss distribution (Arrow, 1963; Pauly, 1974; Rothschild and Stiglitz, 1976; Harris and Raviv, 1978; Holmstrom, 1979; Shavell, 1979). Chiappori et al. (2006) show that this prediction can be extended to general settings, including, for example, heterogeneous preferences and

Electronic copy available at: https://ssrn.com/abstract=2850537

multidimensional hidden information linked with hidden action. However, they point out that the prediction about the positive relation between the level of insurance coverage and risk might no longer hold if the degree of risk aversion in combination with risk type is private information.

The role of private information and its potential effects are adressed by Finkelstein and Poterba (2014) who propose the following extension of Chiappori and Salanié (2000):

$$\text{Coverage} = 1(X\beta_c + Y\gamma_c + \epsilon_c > 0) \tag{4}$$

$$\text{Risk} = 1(X\beta_r + Y\gamma_r + \epsilon_r > 0). \tag{5}$$

Y includes information which is observed or observable but not used by the insurance company, hence Finkelstein and Poterba (2014) call the test “unused observables test.” Under the null hypothesis that there is no private information contained in Y that is relevant for contract choice and risk (i.e. symmetric information), we have $\gamma_c = 0$ and $\gamma_r = 0$. The benefit of this model extension is that the rejection of the null hypothesis directly provides evidence of relevant private information independent of the type of asymmetric information. In our context the information Y (the telematic data) is not observed (and thus unused) by the insurance company but accessible to the econometrician.

This model specification also has implications for the residual correlation test. Finkelstein and Poterba (2014) argue and show that the positive correlation test may fail to reject the symmetric information hypothesis even in the presence of private information about risk type, if there is unobserved heterogeneity in individual preferences. For example, when risk type is positively correlated with both coverage and risk of loss, but risk aversion is positively correlated with coverage and negatively correlated with risk of loss, the correlation between ϵ_c and ϵ_r may be zero or even negative.

Electronic copy available at: https://ssrn.com/abstract=2850537

Unlike in Chiappori and Salanié (2000) and all other literature on car insurance which only considers first-party coverage, policyholders in our data set simultaneously choose the level of coverage along two dimensions, first-party and third-party liability coverage. We define the dependent variables of the outcome equations as follows:

- First-party coverage (Cov1stP): We set Cov1stP $= 1$ if the contract covers at-fault losses (full comprehensive insurance) and Cov1stP $= 0$ otherwise.
- Third-party liability coverage (Cov3rdP): We set Cov3rdP $= 1$ if the upper limit of third-party liability coverage is high and Cov3rdP $= 0$ if the upper limit is low.
- ex-post risk of loss (BMDG): We use downgrades of the Bonus-Malus record to proxy for ex-post risk. The dependent variable BMDG is set to 1 if the Bonus-Malus rating of a policyholder was downgraded within the subsequent year and is set to 0 otherwise.

We specify the following three (outcome) equations:

$$\text{Cov3rdP} = 1(X\beta_3 + Y\gamma_3 + \varepsilon_3 > 0) \tag{6}$$

$$\text{Cov1stP} = 1(X\beta_1 + Y\gamma_1 + \varepsilon_1 > 0) \tag{7}$$

$$\text{BMDG} = 1(X\beta_r + Y\gamma_r + \varepsilon_r > 0) \tag{8}$$

X comprises the set of car- and policyholder-related variables which we consider to determine contract choice and ex-post risk. X includes all relevant variables used by the insurance company for pricing the contract, in particular the distance driven. X also includes weather data[25] as a potential determinant of the driver's choice of coverage and risk of loss. We are not aware that any other study has used weather information as a potential determinant of contract choice and/or risk. A similar effect can be expected from using postcode information (as done by Saito, 2006). However, weather is measured on a metric scale, and is thus much

[25] Further details on weather data will be provided below.

Electronic copy available at: https://ssrn.com/abstract=2850537

better suited to represent location-related information than the nominal postcode scale.

Y includes the private information which is not observed by the insurance company. We test the null hypothesis that there is no private information contained in Y that is relevant for contract choice and risk, i.e. we test for $\gamma_3 = 0$, $\gamma_1 = 0$ and/or $\gamma_r = 0$. We then compare the direct evidence about the relevance of private information with the results obtained from the residual correlation test suggested by Chiappori and Salanié (2000). For that purpose we use generalized residuals ε_3, ε_1, and ε_r to compute the correlation coefficients $\rho_{3,r}$ and $\rho_{1,r}$, both excluding and including the private information Y.[26] Subsequently, we assess whether the conclusions from the results of the residual correlation test are consistent with the direct evidence we derive from the coefficients in equations (6)–(8).

3.2 Endogeneity and sample selection

The unused observables test by Finkelstein and Poterba (2014) has obvious advantages. The private (unused and unobserved) information enhances tests for adverse versus advantageous selection compared to the standard positive correlation test suggested by Chiappori and Salanié (2000). However, "*attributes that are both demand related and correlated with risk of loss*" – as Finkelstein and Poterba (2014) put it – may be problematic in regression-based tests. The main econometric challenge with attributes Y satisfying this property is their potential endogeneity. Elements of Y which represent risk type or risk preferences may both *determine* contract choice (coverage) and may be *affected by* contract choice. For example, drivers having chosen higher coverage may tend to drive riskier, while drivers knowing about their poor driving abilities may choose higher coverage. Including endogenous variables in Y leads to inconsistent estimates of γ *and* β, and invalidates conclusions based on (the correlation between) residuals. Accordingly, it may be problematic to use the private information about driving behavior measured via telematic data. Therefore, we treat speeding,

[26] Since we also include an interaction between rides and distance, we cannot clearly separate private from non-private information. Therefore, we treat the interaction term as private, too.

Electronic copy available at: https://ssrn.com/abstract=2850537

the number of car rides, and the distance driven as endogenous regressors.

The standard remedy for endogeneity is using instrumental variables (see, for example, Dionne et al., 2009, 2013). Suitable instruments (a) must be exogeneous, (b) have to be good predictors of the endogeneous regressors, and (c) must not be included as regressors in the equations of interest. One can think of instrumental-variable (IV) estimation as a two-step procedure (done by standard estimation methods in a single step). In the first-stage, the potentially endogeneous regressor $y \in Y$ is regressed on all other (exogeneous) regressors and the instruments. Subsequently, in the equation of interest (the outcome equation), the endogenous y is replaced by fitted values $\hat{y}$ from the first-stage regression.

We address the endogeneity problem by using local, car-related weather conditions as instruments. Weather conditions observed after the coverage choice, during the period when driving information is recorded (i.e. February to April 2009), are suitable instruments because (a) weather is exogenous, (b) it affects driving behavior, and (c) weather observed during the three months satisfies the exclusion condition, since it is not relevant for coverage choice and downgrades.[27]

More specifically we use data on precipitation (rainfall), snowfall (depth and number of days), temperature at 7am and 2pm, visibility (number of foggy and hazy days), and radiation. We use the average, the minimum, the maximum, and the range (i.e., the difference between maximum and minimum) over the three months, resulting in a total of 36 measures. We obtain this data from about 150 weather stations and match it with the main location of the car. Matching is done by using data from the weather station with the shortest distance to the main location. To simplify the search for an appropriate set of instruments, we do not consider the originally observed weather data but derive and use the three principal components with eigenvalues greater than one, which are significant predictors in the first-stage (see the results below). It is difficult to summarize the structure of loadings for the three components. The first factor is primarily associated with temperature and the second

[27] We discuss the exclusion condition more comprehensively below.

Electronic copy available at: https://ssrn.com/abstract=2850537

with visibility. For the third we find no clear pattern in loadings allowing such associations. In any case, by using these three components, we efficiently combine information from a large number of weather measures and also avoid the problem of using too many instruments.

As noted above, the four measures which characterize driving behavior (speeding, number of car rides, distance driven, and the interaction between rides and distance) are treated as endogenous. Accounting for multiple endogenous regressors leads to special requirements. The usual concerns associated with weak instruments are exacerbated, and the multivariate case has to be treated differently (see Stock et al., 2002). It is necessary to find instruments which can sufficiently differentiate among two or more endogeneous regressors so that the fitted values from the first-stage regressions are not highly correlated. As it turns out, it is very difficult to define instruments with that capability using the available weather data. Instead, we combine the four endogeneous driving-related measures into a *single* endogeneous variable, defined as their first principal component. This component has an eigenvalue of 2.7, and the loadings of distance driven, the number of car rides and their interaction are between 0.5 and 0.6. The loading of the speeding index is 0.2. We interpret this principal component as a driving-based, endogenous risk factor. In the subsequent analysis we compare results based on using only the distance driven (together with all other controls) to results based on this driving-factor (which includes both private driving measures and the non-private distance). Although we forgo some interpretability we avoid the challenge of multiple endogenous regressors and the associated problem of weak instruments.

For the estimation of the three outcome equations (6), (7), and (8) in combination with sample selection we follow Wooldridge (2010), section 19.6.[28] We first estimate a selection equation using data on a sample of 1987 clients who did not choose the telematic contract (with the binary dependent variable being zero), and the 1847 clients who opted for the telematic contract. The inverse-Mills ratio derived from this probit regression is used in

[28] We note that Wooldridge (2010) does not address our case of endogeneity in the *binary* response equation and remarks that this “is difficult, and it is a useful area of future research” (p.814). He also does not account for the case of more than one outcome equation.

Electronic copy available at: https://ssrn.com/abstract=2850537

all subsequent regressions to avoid a potential sample-selection bias. We estimate the three outcome equations as separate[29] IV-probit models, defining Y in (6)–(8) as the first principal component of the four driving-based measures.

In the estimation of the selection equation, and the (implicit) first-stage and outcome equations, we use three different kinds of weather information for three different purposes:

- Long-run weather data from 1981 to 2010 is used in the two coverage choice equations (6) and (7), and weather from 2009 in the BM downgrade equation (8). It is plausible to assume that policyholders are aware of typical (local) weather conditions near the main location. Accordingly, they may account for such circumstances in their coverage choice. More specifically, we use long-run precipitation matched with the telematic-based main location of a car in the coverage choice equations. Differences in local weather conditions may also be responsible for differences in the likelihood of damages and consequential downgrades. However, it is more appropriate to use weather conditions during 2009 rather than long-run weather data in the BM downgrade equation, since downgrades reflect rating changes during 2009. Therefore, we use the first principal component of 36 weather variables measured in the entire year 2009 in the BM downgrade equation (8).

- Weather observed during the three months (February, March and April 2009) while telematic data is recorded is used to derive three principal components. As described above, these are used as instruments to account for the endogeneity of the driving-factor in the IV-probit estimation. Instruments have to satisfy the exclusion condition, and thus must not be included in the outcome equations (6)–(8). We cannot rule out that (anticipated) weather conditions during these three months may have some relevance for coverage choice, and may have a partial impact on downgrades. However, we argue

[29] Estimating the three probit equations jointly would imply efficiency gains (which may not be substantial given the weak correlations among residuals presented below). However, simultaneously accounting for endogeneity in the joint estimation is not straightforward. Rather than obtaining potentially inconsistent results, we prefer to present single-equation estimates which may be (slightly) inefficient.

Electronic copy available at: https://ssrn.com/abstract=2850537

that *long-run* weather conditions (represented by long-run precipitation) for coverage choice and weather conditions for the *entire year* 2009 for downgrades (see the previous bullet point) are *better suited* as controls in the outcome equations than the weather observed only during these three months.[30]

- The selection equation requires an exogeneous regressor which is not included in subsequent equations. Ideally, such a factor should determine the selection of the pay-as-you-drive contract but should not affect coverage choice and BM downgrades (see Wooldridge (2010), 19.6.2). We use the first[31] principal component of long-run weather measures (e.g. precipitation, temperature, the number of days with snow, hail or ice) observed over the period 1981-2010, assuming that this information reflects (exogenous) regional differences which affect sample selection.[32] The choice of the selection instrument is not as critical as in IV-estimation. If it is too weak, the inverse Mills ratio derived from the selection equation may be too strongly correlated with the other regressors subsequently used in the (implicit) first stage of IV-estimation. However, this correlation is attenuated for three reasons: (a) only (about) half of the sample used in the selection equation is also used in the subsequent regressions, (b) the inverse Mills ratio is a nonlinear function of fitted values, and (c), as shown below in Table 3, the weather-based selection instrument has a significant coefficient in the selection equation.

[30] The correlation of the first of the three '3-month weather' principal components with long-run precipitation is 0.66 for the subsample of policyholders who chose the pay-as-you-drive contract. Its correlation with the 'weather 2009' component is 0.77. These correlations do not seem too high, given the conceptual differences between those measures.

[31] The first component has an eigenvalue of 3.9. All other components have eigenvalues less than 1.0.

[32] In the selection equation we use this principal component as an instrument while we use long-run precipitation in the outcome equations (6) and (7) as a determinant of coverage choice. The correlation between long-run precipitation and the first principal component of long-run weather is 0.58 for all data (compared to 0.66 for the subsample of policyholders who chose the pay-as-you-drive contract). The correlation between long-run precipitation and the first of the three '3-month weather' components is 0.73 using all data.

Electronic copy available at: https://ssrn.com/abstract=2850537

Table 3: Selection equations

	Bonus-Malus downgrade		first- and third party coverage	
	coef	z-stats	coef	z-stats
log HP	-0.663	-6.2	-0.670	-6.3
log adjusted price	1.119	17.8	1.104	17.7
log age car	0.866	14.6	0.859	14.6
gender (male=1)	-0.153	-3.0	-0.161	-3.1
age	-0.016	-9.3	-0.016	-9.5
BM rating	2.890	6.9	2.932	7.0
impact downgrade	5.021	8.0	4.851	7.8
impact upgrade	-1.274	-1.0	-1.203	-1.0
urban	0.370	6.3	0.436	7.4
log purchasing power	0.976	5.1	0.906	4.7

Notes: Probit regression; dependent variable =1 if pay-as-you-drive constract has been chosen; columns 2 and 3 report results for the equation used to compute the inverse Mills ratio for the BMDG equation; columns 4 and 5 for the coverage outcome equations; N=3834; pseudo R^2: 0.27 and 0.28; robust standard errors are used to compute z-statistics.

4 Results and Discussion

4.1 Results

The pay-as-you-drive insurance contract is offered for choice, and the characteristics of policyholders under this contract might differ from those that decided not to choose this contract. The results for the selection equation[33] presented in Table 3 show that the pay-as-you-drive contract is more likely to be chosen by younger, female policyholders living in urban and/or wealthier areas, who drive old(er) and/or more valuable cars with less engine power. The pay-as-you-drive contract is also more likely to be chosen by policyholders whose current Bonus-Malus rating is high (indicating high risk), and whose premium is more strongly affected by a potential downgrade. A potential upgrade has no significant effect.

In Table 4 we summarize results related to weather variables. Long-run weather, repre-

[33] Because we use different controls in the outcome equations for coverage and for BM downgrades we estimate two, slightly different, selection equations.

Electronic copy available at: https://ssrn.com/abstract=2850537

Table 4: z-statistics of weather variables in selection equations and first-stage IV regressions

			first-stage of IV-estimation			
	selection eqn		distance		driving factor	
	BMDG	coverage	BMDG	coverage	BMDG	coverage
long-run weather pc	-2.3	-4.3				
3-month weather pc1	7.0	3.8	-5.3	-5.1	-4.1	-3.5
3-month weather pc2	6.4	4.3	1.4	2.4	1.5	2.6
3-month weather pc3	6.8	7.2	-4.1	-4.0	-6.0	-5.8
long-run precipitation		-4.5		-0.4		1.2
weather 2009 pc	-8.3		1.4		2.4	

Notes: robust z-statistics associated with weather variables in the selection equations (binary probit with the same dependent variable as in Table 3), and the first-stage IV regressions (dependent variables are distance and the driving-factor); we only present z-statistics because coefficients of principal components have no meaning; results for other regressors can be found in Table 3, or have been omitted for brevity.

sented by the first principal component of various long-run weather measures, is found to be a significant regressor in the selection equations, justifying its use as an instrument for this purpose.[34] In Table 4 we only present the weather-related part of first-stage instrumental variable estimation results. For the treatment of endogeneity, it is important that the coefficients of 3-month weather instruments are significant. The second principal component is insignificant, but only in the first-stage of the BM downgrade equation. Overall, however, the instruments can be considered to be strong, given their F-statistics are above 25 (i.e. above the typical threshold 10, see Stock et al. 2002), and partial R^2's are around 0.04.

We now turn to the main results of our study. Table 5 presents the estimates of the three outcome questions. For each binary dependent variable, we present two regression results. The first regression includes only distance as a driving-based regressor that the insurance company observes and includes in the pricing scheme. In the second regression, we include the driving-factor that summarizes information about the distance driven, the number of

[34] The principal component of long-run weather is not used in subsequent regressions to satisfy the exclusion condition. The other weather-related regressors have to be included in the selection equation because they are subsequently used in the outcome equations and as instruments for treating endogeneity.

Electronic copy available at: https://ssrn.com/abstract=2850537

car rides, the interaction between these two, and the speeding index.[35] This driving-factor is based on information that is private to the car driver (and to us), but also includes distance, the non-private information observed and used by the insurance company. We treat endogeneity of both driving-based regressors, distance and the driving-factor, with three instruments.

From column ‘BM downgrade’ in Table 5 we infer that both distance and the driving-factor have a positive effect on a downgrade of the BM rating class. While the signs of the coefficients are suggestive of distance and driving-factor being risk factors, the z-statistics of 1.0 and 1.3, respectively, are close to but below the usual critical values.

Regarding the effect of the two driving-based regressors on insurance coverage (see columns ‘3rd party coverage’ and ‘1st party coverage’), the signs indicate a negative effect on third-party liability coverage (indicating advantageous selection) and a positive effect on first-party coverage (indicating adverse selection). However, the coefficients are all insignificant (the z-statistics are even lower than the ones for ‘BM downgrade’). In summary, the effects of the driving-factor (based on private information) on BM downgrade and coverage choice do not provide empirical support for either advantageous or adverse selection.

We now turn to analyzing the correlation among generalized residuals. Table 6 presents correlations and associated p-values between coverage choice (both third-party liability and first-party) and risk (BM downgrade) for three models. The first model (1), is a baseline model that serves as a benchmark. We derive the generalized residuals in this model from a probit model that only includes non-driving based information in each outcome equation, i.e., neither distance nor the driving-factor is included. In the second model (2) and in third model (3), we then add distance and the driving-factor, respectively, as regressors to the baseline model and derive the generalized residuals accordingly.

The results of the baseline model (1) show that the correlations between coverage and risk are

[35] According to a standard setting of principal components analysis the variance of each component is equal to its eigenvalue. We scale the driving-factor such that it has the same variance as distance, thus making its coefficients in Table 5 comparable to those of distance in the first row.

Electronic copy available at: https://ssrn.com/abstract=2850537

Table 5: Outcome equations

	BM downgrade				3rd party coverage				1st party coverage			
	coef	z	coef	z	coef	z	coef	z	coef	z	coef	z
distance	0.356	.96			-0.148	-.70			0.050	.17		
driving factor			0.435	1.3			-0.168	-.85			0.036	.13
log HP	-0.376	-1.9	-0.345	-1.7	0.098	0.7	0.086	0.6	-0.572	-2.6	-0.569	-2.6
log adjusted price	0.358	1.8	0.351	1.9	-0.041	-0.3	-0.038	-0.3	0.690	3.4	0.694	3.5
log age car	0.693	4.1	0.678	4.0	-0.201	-1.8	-0.198	-1.8	-0.732	-4.5	-0.732	-4.5
gender (male=1)	-0.070	-0.7	-0.078	-0.8	0.038	0.5	0.040	0.6	-0.004	0.0	-0.002	0.0
age	0.005	0.6	0.006	0.9	0.004	0.7	0.003	0.6	-0.008	-1.1	-0.008	-1.2
BM rating	0.493	1.0	0.439	0.9	0.422	1.0	0.449	1.1	-0.218	-0.5	-0.219	-0.5
impact downgrade	1.572	1.5	1.670	1.7	-2.574	-3.4	-2.594	-3.4	-1.115	-1.2	-1.111	-1.2
impact upgrade	0.593	0.4	0.382	0.3	1.778	1.4	1.884	1.5	0.655	0.5	0.603	0.5
urban	0.174	1.3	0.198	1.5	-0.120	-1.4	-0.129	-1.5	0.263	2.5	0.261	2.4
log purch. pow.	0.144	0.5	0.207	0.7	-0.159	-0.7	-0.182	-0.8	0.717	2.3	0.722	2.3
long-run prec.					0.011	2.6	0.012	3.2	0.012	2.3	0.011	2.6
weather 2009 pc	-0.019	-0.9	-0.026	-1.6								
inverse Mills ratio	0.406	1.9	0.397	1.8	-0.241	-1.4	-0.232	-1.3	0.319	1.3	0.321	1.3

Notes: IV-probit regressions; N=1847; robust standard errors are used to compute z-statistics; 'driving-factor' denotes the first principal component of the endogeneous regressors (i.e. the number of car rides, the distance driven, the interaction among these two, and the speeding index); three principal components of weather measured during the three months of the observation period are used as instruments.

not statistically significantly different from zero.[36] Following Chiappori and Salanié (2000), we would interpret these results as evidence for the absence of private and relevant information.

The fact that this result does not change when adding distance (model (2)) and the driving-factor (model (3)) as regressors is consistent with our results in Table 5. We can interpret the correlations between the generalized residuals of model (2) and (3) as measures for the link between coverage choice and risk, after non-private information about distance and private driving-based information has been removed. However, neither distance nor the driving-factor show a statistically significant effect on risk and on coverage choice. Therefore, the signs and significance of the correlations among the generalized residuals should not and do

[36] The signs suggest a negative correlation between third-party liability coverage and risk and a positive correlation between first-party coverage and risk.

Electronic copy available at: https://ssrn.com/abstract=2850537

Table 6: Correlations among generalized residuals

	(1) w/o driving information		(2) only distance added		(3) driving-factor added	
	ρ	p-value	ρ	p-value	ρ	p-value
Cov3rdP and BMDG	-0.001	0.952	-0.014	0.543	-0.015	0.526
Cov1stP and BMDG	0.036	0.112	0.034	0.120	0.035	0.108

Notes: Correlations are based on generalized residuals derived from the outcome equations in Table 5 (except column (1)). We report the associated p-value of the χ^2 test statistic W used by Chiappori and Salanié(2000).

not change when adding distance or the driving-factor as regressors.

In addition to testing for asymmetric information based on residual correlations we have also applied the test suggested by Dionne et al. (2013), i.e., we have estimated their equation (1) for both first- and third-party coverage choice. The coefficient of BM downgrades (γ in their paper) is insignificantly negative for Cov3rdP (p-value 0.9) and insignificantly positive for Cov1stP (p-value 0.2). Thus, signs and significance correspond to the evidence provided in Table 6, and also indicate the absence of residual asymmetric information.

We finally comment on some further results, which are not directly related to the role of private information, adverse/advantegeous selection and moral hazard. Most of the effects of controls in Table 5 have the expected signs, and some are highly significant (notably, the age of cars in the BM downgrade equation). The higher the increase in the premium in case of (a potential future) downgrade, the less third-party coverage is demanded. The latter effect can be interpreted as the result of financial considerations. We find significant weather-related effects only in the coverage-choice equations; the demand for first- and third-party coverage increases with long-run precipitation. The coefficients associated with engine power, the price of a car, and age in the first-party choice equation are highly significant, indicating that this choice is primarily driven by these factors.[37]

[37] The rather large coefficients reflect the similar role of these regressors, and are not due to strong multicollinearity. The correlation between the log of adjusted price and the log of car age is −0.76, the correlation between log of HP and log of adjusted price is 0.53. The negative coefficient of engine power

Electronic copy available at: https://ssrn.com/abstract=2850537

Purchasing power has the expected positive and significant coefficient in first-party coverage choice. A better measure for income (with more variation than purchasing power, see Table 2), and/or a better match with the policyholder would be preferable. This could potentially make a difference to the insignificant negative effect in the third-party coverage equation. It may also be the case that other regressors measured on an individual level, for instance, the impact of a BM downgrade, are picking up financial aspects in third-party coverage choice better than purchasing power.

We finally note that the coefficient of the inverse Mills ratio is significantly positive (with a p-value of $\approx$0.05) in the BM downgrade equation, but not in the coverage choice equations. This indicates that policyholders which are more likely to be downgraded tend to have chosen the pay-as-you-drive contract, while there are no sample selection effects with respect to coverage choice.

4.2 Robustness checks

As discussed in the section on endogeneity, the distinction between non-private and private driving-related information is a conceptually important yet empirically subtle point. Therefore, we combine all (endogeneous) driving-based measures into a single factor. In order to obtain some insights into the relative importance of its components, we also consider other combinations of the driving measures.

We obtain the factor 'distance&rides' by combining distance (which has always to be included since it is non-private information observed and used by the insurance company) with the number of rides and the interaction term (each with about the same loading). We also combine distance with the speeding index to obtain the factor 'distance&speeding' (using the same loadings for both variables).[38] Row 'distance&rides factor' in Table 7 indicates that

may reflect the effects of different car brands.

[38] We again scale the driving-factors such that they have the same variance as distance to make the coefficients of driving-factors in Table 7 comparable to those of distance in the first row.

Electronic copy available at: https://ssrn.com/abstract=2850537

Table 7: Estimates for alternative driving-based factors, but using alternative methods to define instruments

	BMDG		Cov3rdP		Cov1stP	
	coef	z	coef	z	coef	z
distance	0.356	0.96	-0.148	-0.70	0.050	0.17
driving-factor	0.435	1.34	-0.168	-0.85	0.036	0.13
distance&rides factor	0.491	1.65	-0.175	-0.89	0.038	0.14
distance&speeding factor	0.309	0.44	-0.191	-0.64	0.000	0.00
distance&placebo factor	0.408	0.95	-0.172	-0.72	0.052	0.15
driving-factor using						
stepwise	0.525	(2.0)	-0.098	-(1.2)	-0.045	-(0.4)
lasso	0.445	(1.4)	-0.134	-(1.6)	0.057	(0.5)

Notes: The rows 'distance' and 'driving-factor' are copied from Table 5. The estimates in row 'distance&placebo factor' are averages from 200 simulation runs. This factor consists of observed distance and a placebo private driving measure which is uncorrelated with all outcome variables. The estimates in the rows 'stepwise' and 'lasso' are based on a two-stage instrumental variable procedure which does not provide correct standard errors. Therefore z-statistics are too large (in absolute terms) and put in parentheses.

excluding speeding from the set of endogenous driving measures results in a larger coefficient and z-statistic of this factor compared to the driving-factor in the BM downgrade equation. The results in the other two equations remain unaffected. The attenuating role of speeding is confirmed in the row 'distance&speeding factor' which shows that this driving-based factor is not significant at all.

One may be concerned that the role of the non-private measure 'distance' in the driving-factor is too strong to find something potentially important about the private information. The correlation between the principal component score (i.e. the driving-factor) and the observed distance is 0.9. The loading of distance on this principal component is 0.5, which implies that the relative weight of distance in the linear combination constituting the driving-factor is 0.3. Rather than judging whether or not these values are (too) high or low (enough), we ask whether correlations, loadings, and weights really indicate whether the private information embedded in the driving-factor is sufficiently different from the non-private information to be convincing.

Electronic copy available at: https://ssrn.com/abstract=2850537

To answer this question we carry out a placebo test. We use the same dataset as above, retain all variables, apply the same estimation procedure (accounting for sample selection and endogeneity), and add a simulated standard normal random variable. For the placebo test we construct a driving-factor which consists of distance and the placebo. This distance&placebo-factor is simulated such that it has a correlation of 0.9 with observed distance (i.e. the same correlation as between distance and the driving-factor used above). The artificial variable included in this factor is exogeneous and uncorrelated with BM downgrades and coverage choice. It has no ex-ante (economic) support and should be irrelevant for risk classification. Hence, including such pseudo-private information in the analysis should not change the significance of the driving-factor in the outcome equations. Upon comparing the results in the first two rows of Table 7 (copied from Table 5) to the row 'distance&placebo factor' we find that the z-statistics from the placebo test are very similar to the first row ('distance'), and the coefficients from the placebo test are comparable to the coefficients in the second row ('driving-factor').[39] A similar stability holds for the correlations reported in Table 6.

Hence, a driving-factor which includes irrelevant private information does not make a difference even though it is highly correlated with its non-private constituent 'distance'. In contrast, the results change when considering the driving-factor we construct by adding private information (i.e. number of rides and speeding). Compare the z-statistic of the driving-factor in the BM downgrade equation 1.34 to the z-statistic of distance alone (0.96 in the first row of Table 7). The significance does change upon adding ex-ante relevant private information. Hence, there appears to be sufficient additional private information in the driving-factor which is related to downgrades 'more systematically' than distance alone.[40] This view is supported by considering the results for the factor which is based on only distance and the number of rides. The significance of the distance&rides-factor increases even

[39] Note that we treat the 'distance&placebo factor' in the same way as the (endogenous) driving-factor. However, the placebo component in this factor is exogeneous, and we cannot expect its coefficients to be very similar to those of the driving-factor.

[40] An increase in the significance indicates an improved risk classification ability when additional private information is accounted for.

Electronic copy available at: https://ssrn.com/abstract=2850537

further (the z-statistic in the BM downgrade equation is 1.65). Conversely, constructing a driving-factor by adding speeding to distance results in an insignificant coefficient in the downgrade equation. This indicates rather noisy and unsystematic information conveyed by speeding, which deteriorates the risk classification ability of distance and/or distance and rides. Given the results from the placebo test, we conclude that observed changes in significance are not an artefact of multicollinearity, and there is sufficient potential for the unobserved parts of driving behavior to be important.

So far we have used instruments defined as the principal components of weather measures observed during the three months while driving was observed. In addition, we have also considered other methods which all aim at extracting predictive information from a large number of regressors while reducing the risk of overfitting. We consider a stepwise-selection procedure, and the lasso-approach suggested by Tibshirani (1996). The last two rows in Table 7 which correspond to these two approaches confirm the results obtained in Table 5. The driving-factor has a (weak) positive effect on BM downgrades, a weak(er) negative effect on third-party coverage, and no effect on first-party coverage.

The first principal component of weather observed in 2009 has been considered as a potentially relevant factor for BM downgrades. The results in Table 5 indicate, however, that it is not significant. At the same time, we obtain significant coefficients for long-run precipitation in both coverage-choice equations. Therefore, we have also specified BM downgrade equations using precipitation from 2009, and/or other individual measures of weather observed in 2009. However, no such attempt has resulted in significant coefficients for 2009 weather variables in that equation.

4.3 Conclusion and Discussion

We use a unique data set of driving behavior to test whether private information in driving characteristics has significant effects on contract choice and/or risk. The data set is unique in

Electronic copy available at: https://ssrn.com/abstract=2850537

several ways. First, it provides detailed GPS information about the position and movement of cars. Second, with the exception of distance driven, the data is not accessible to the insurance company. The data thus provides us with the opportunity to directly test the relevance of private information that results from car drivers' actions that are relevant to but not observed by the insurance company. Third, the data represents all pay-as-you-drive insurance contracts on the market at that time.

Moreover, we contribute to the literature by applying a methodological framework to account for sample selection effects and endogeneity. We use local weather conditions, matched to the location of cars, as instrumental variables. Thereby, we make it possible to disentangle adverse selection and moral hazard effects.

We define a driving-factor which includes overall distance driven (known to the insurance company), as well as the number of car rides and speeding (private information not accessible by the insurer). We find evidence that *additionally* including private information in a driving-factor affects a downgrade of the Bonus-Malus class more strongly than non-private distance alone. At the same time none of the driving characteristics affect policyholders' contract choice.

In their survey paper, Cohen and Siegelman (2010) list several reasons that might explain the absence of selection effects despite the existence of private information. For example, policyholders might not be aware of their informational advantage, that is, they do not know how their driving behavior affects accident risk and thus do not sort into different contracts.

In the context of driving and insurance purchase decisions, an additional explanation might be that individuals' decisions regarding financial risk taking (purchasing insurance to avoid losses) need not be systematically related to their decisions regarding non-financial risk taking (driving carefully to avoid accidents).

Salminen and Heiskoanen (1997) present evidence that there are differences in the degree to which individuals are exposed to different types of non-financial risks, such as traffic, occu-

Electronic copy available at: https://ssrn.com/abstract=2850537

pational, or home accidents. Moreover, Einav et al. (2012) show that there are differences in risk preferences for financial risks across different domains, such as insurance and asset allocation. Thus, there might be even larger differences in how individuals deal with financial (insurance) versus non-financial (automobile) risks. Some individuals might be much more willing to avoid non-financial risks by driving very carefully while they are less willing to insure financial risks by purchasing automobile insurance, and vice-versa. Consistent with our results, driving behavior might therefore have little relevance for the choice of insurance contracts.

Electronic copy available at: https://ssrn.com/abstract=2850537

References

[1] Arrow, K.J., 1963, Uncertainty and the Welfare Economics of Medical Care, *American Economic Review* 53(5): 941-973

[2] Bucciol, A., and R. Miniaci, 2011, Household Portfolios and Implicit Risk Preference, *Review of Economics and Statistics* 93(4): 1235-1250

[3] Cawley, J., and T. Philipson, 1999, An Empirical Examination of Information Barriers to Trade in Insurance, *American Economic Review* 89(4): 827-846

[4] Chiappori, P.-A., B. Jullien, B. Salanié, and F. Salanié, 2006, Asymmetric Information in Insurance: General Testable Implications, *RAND Journal of Economics* 37(4): 783-798

[5] Chiappori, P.-A., and B. Salanié, 2000, Testing for Asymmetric Information in Insurance Markets, *Journal of Political Economy* 108(1): 56-78

[6] Cohen, A., 2005, Asymmetric Information and Learning in the Automobile Insurance Market, *Review of Economics and Statistics* 87(2): 197-207

[7] Cohen, A., and L. Einav, 2007, Estimating Risk Preferences from Deductible Choice, *American Economic Review* 97(3): 745-788

[8] Cohen, A., and P. Siegelmann, 2010, Testing for Adverse Selection in Insurance Markets, *Journal of Risk and Insurance* 77(1): 39-84

[9] Cutler, D.M., and S.J. Reber, 1998, Paying for Health Insurance: The Trade-Off Between Competition and Adverse Selection, *Quarterly Journal of Economics* 113(2): 433-466

[10] Cutler, D.M., and R.J. Zeckhauser, 1998, Adverse Selection in Health Insurance, *Forum for Health Economics and Policy*: Vol. 1: (Frontiers in Health Policy Research): Article 2. http://www.bepress.com/fhep/1/2

Electronic copy available at: https://ssrn.com/abstract=2850537

[11] de Meza, D., and D.C. Webb, 2001, Advantageous Selection in Insurance Markets, *RAND Journal of Economics* 32(2): 249-262

[12] Dionne, G., and L. Eeckhoudt, 1985, Self-Insurance, Self-Protection and Increased Risk Aversion, *Economics Letters* 17(1-2): 39-42

[13] 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* 109(2): 444-453

[14] Dionne, G., P. St-Amour, D. Vencatachellum, 2009, Asymmetric Information and Adverse Selection in Mauritian Slave Auctions, *Review of Economic Studies*, 76: 1269-1295

[15] Dionne, G., P-C. Michaud, J. Pinquet, 2013, A Review of Recent Theoretical and Empirical Analyses of Asymmetric Information in Road Safety and Automobile Insurance, *Research in Transportation Economics* 43: 85-97

[16] Edlin, A.S., 2003, Per-Mile Premiums for Auto Insurance, in *Economics for an Imperfect World: Essays In Honor of Joseph Stiglitz*, Ed. Richard Arnott, Bruce Greenwald, Ravi Kanbur, Barry Nalebuff, MIT Press, 53-82

[17] Edlin, A.S., and P. Karaca-Mandic, 2006, The Accident Externality from Driving, *Journal of Political Economy* 114(5): 931-955

[18] Ehrlich, I, and G. Becker, 1972, Market Insurance, Self-Insurance, and Self-Protection, *Journal of Political Economy* 80(4): 623-648

[19] Einav, L., A. Finkelstein, I. Pascu, and M.R. Cullen, 2012, How General Are Risk Preferences? Choices under Uncertainty in Different Domains, *American Economic Review* 102(6): 2606-2638

[20] The European Commission, 2011, Commission Recommendation of 8 September 2011 on support for an EU-wide eCall service in electronic communication networks for the transmission of in-vehicle emergency calls based on 112 ('eCalls'): *Official Journal of the European Union* L 303, 22.11.2011, 46-48

Electronic copy available at: https://ssrn.com/abstract=2850537

[21] Fang, H., M.P. Keane, and D. Silverman, 2008, Sources of Advantageous Selection: Evidence From the Medigap Insurance Market, *Journal of Political Economy* 116(2): 303-350

[22] Finkelstein, A., and K. McGarry, 2006, Multiple Dimensions of Private Information: Evidence From the Long-Term Care Insurance Market, *American Economic Review* 96(4): 938-958

[23] Finkelstein, A., and J. Poterba, 2004, Adverse Selection in Insurance Markets: Policy-holder Evidence From the U.K. Annuity Market, *Journal of Political Economy* 112(1): 183-208

[24] Finkelstein, A., and J. Poterba, 2014, Testing for Asymmetric Information using 'Unused Observables' in Insurance Markets: Evidence from the U.K. Annuity Market, *Journal of Risk and Insurance* 81(4): 709-734

[25] Gan, L., M.D. Hurd, and D.L. McFadden, 2005, Individual Subjective Survival Curves, in D. Wise (ed.), *Analyses in the Economics of Aging*, Chicago: University of Chicago Press, 377-411

[26] Harris, M., and A. Raviv, 1978, Some Results on Incentive Contracts With Applications to Education and Employment, Health Insurance, and Law Enforcement, *American Economic Review* 68(1): 20-30

[27] Hendren, N., 2013, Private Information and Insurance Rejections, *Econometrica*, 81(5): 1713-1762

[28] Holmstrom, B., 1979, Moral Hazard and Observability, *Bell Journal of Economics* 10(1): 74-91

[29] Hurd, M.D., 1999, Anchoring and Acquiescence Bias in Measuring Assets in Household Surveys, *Journal of Risk and Uncertainty* 19(1-3): 111-136

[30] Hurd, M.D., D.L. McFadden, H. Chand, L. Gan, A. Merrill, and M. Roberts, 1998, Consumption and Saving Balances of the Elderly: Experimental Evidence on Survey

Electronic copy available at: https://ssrn.com/abstract=2850537

Response Bias, in D. Wise (ed.), *Topics in the Economics of Aging*, Chicago: University of Chicago Press, 353-87

[31] Insurance Research Council, Uninsured Motorists and Public Attitude Monitoring, 2000 - 2003

[32] Jullien, B., S. Salanié, and F. Salanié, 1999, Should More Risk-Averse Agents Exert More Effort?, *The Geneva Papers on Risk and Insurance Theory* 24(1): 19-28

[33] Jullien, B., S. Salanié, and F. Salanié, 2007, Screening Risk-Averse Agents Under Moral Hazard: Single-Crossing and the CARA Case, *Economic Theory* 30(1): 151-169

[34] Koufopoulos, K., 2007, On the Positive Correlation Property in Competitive Insurance Markets, *Journal of Mathematical Economics* 43(5): 597-605

[35] Koufopoulos, K., 2008, Asymmetric Information, Heterogeneity in Risk Perceptions and Insurance: An Explanation to a Puzzle, SSRN Working Paper 1301522

[36] McCarthy, D., and O.S. Mitchell, 2010, International Adverse Selection in Life Insurance and Annuities, in S. Tuljapurkar, N. Ogawa, and A.H. Gauthier (eds.), *Ageing in Advanced Industrial States: Riding the Age Waves*, Vol. 3, Springer, 119-135

[37] Morin, R., and A. Suarez, 1983, Risk Aversion Revisited, *Journal of Finance* 38(4): 1201-1216

[38] Parry, I.W.H., 2005, Is Pay-as-you-drive Insurance a Better Way to Reduce Gasoline than Gasoline Taxes?, *American Economic Review: Papers and Proceedings* 95(2): 288-293

[39] Pauly, M.V., 1974, Overinsurance and Public Provision of Insurance: The Role of Moral Hazard and Adverse Selection, *Quarterly Journal of Economics* 88(1): 44-62

[40] Puelz, R., and A. Snow, 1994, Evidence on Adverse Selection: Equilibrium Signaling and Cross-Subsidization in the Insurance Market, *Journal of Political Economy* 102(2): 236-257

Electronic copy available at: https://ssrn.com/abstract=2850537

[41] Robinson, P.A., F.A. Sloan, and L.M. Eldred, 2018, Advantageous Selection, Moral Hazard, and Insurer Sorting on Risk in the U.S. Automobile Insurance Market, *The Journal of Risk and Insurance* 85(2): 545-575

[42] Rothschild, M., and J. Stiglitz, 1976, Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information, *Quarterly Journal of Economics* 90(4): 629-649

[43] Saito, K., 2006, Testing for Asymmetric Information in the Automobile Insurance Market Under Rate Regulation, *Journal of Risk and Insurance* 73(2): 335-356

[44] Salanié, B., 2017, Equilibrium in Insurance Markets: An Empiricist's View, *The Geneva Risk and Insurance Review* 42(1): 1-14

[45] Salminen, S. and M. Heiskanen, 1997, Correlations between traffic, occupational, sports, and home accidents, *Accident Analysis & Prevention* 29(1): 33-36

[46] Sandroni, A., and F. Squintani, 2013, Overconfidence and asymmetric information: The case of insurance, *Journal of Economic Behavior & Organization* 93, 149-165

[47] Shavell, S., 1979, On Moral Hazard and Insurance, *Quarterly Journal of Economics* 93(4): 541-562

[48] Stock, J., J.H. Wright, and M. Yogo, 2002, Weak Instruments and Identification in GMM, *Journal of Business and Economic Statistics* 20(4): 518-529

[49] Tibshirani, R., 1996, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society (Series B)*, 58(1): 267-288

[50] Vickrey, W., 1968, Automobile Accidents, Tort Law, Externalities and Insurance: An Economist's Critique, *Law and Contemporary Problems* 33(3): 464-487

[51] Wooldridge, J., 2010, Econometric Analysis of Cross Section and Panel Data, 2nd edition, MIT Press.

Electronic copy available at: https://ssrn.com/abstract=2850537