Institute of Insurance Economics

University of St.Gallen

THE IDENTIFICATION OF INSURANCE FRAUD – AN EMPIRICAL ANALYSIS –

KATJA MÜLLER

WORKING PAPERS ON RISK MANAGEMENT AND INSURANCE NO. 137

EDITED BY HATO SCHMEISER

CHAIR FOR RISK MANAGEMENT AND INSURANCE

JUNE 2013

The Identification of Insurance Fraud – An Empirical Analysis –

Katja Müller*

Abstract

Fraud is a major concern in the insurance industry. Time after time, spectacular incidents become public of individuals trying to scam tremendous indemnifications from their insurance companies. The majority of claims, however, particularly those seeking low to medium indemnification, exhibit no obvious signs of fraudulent activity thereby leading the insurer to believe they were legitimate. In this study, we therefore focus on determining the characteristics that make an accurate distinction between fraudulent and legitimate claims possible. In addition to identifying dishonest cases more systematically, applying a criteria catalog would enable an efficient use of the limited resources with which fraud investigation divisions are usually endowed. The basis of our analysis is established by a comprehensive data set of automobile claims from a large Swiss insurance company collected throughout the years of 2004 to 2011. The results of the logistic regression analyses reveal different relevant determinants on the policyholder, vehicle, policy and loss level. Contrary to common assumptions, it is most often individuals with a flawless driving record possessing new, high-valued cars who decide to defraud their insurance company. In extension, we place special focus on how the amount of loss affects an individual's likelihood of engaging in fraudulent activities.

*The author is with the Institute of Insurance Economics, University of St. Gallen, Tannenstr. 19, 9000 St. Gallen, Switzerland.

1 Introduction

Insurance fraud has been a key concern in the industry ever since. To date, particularly astonishing incidents have regularly made headlines involving tremendous illegitimate indemnifications from insurance companies. These cases, however, may just be the tip of the iceberg. According to a report by the Association of British Insurers (2012), 15 fraud attempts are being detected each hour of every day, summing up to 139,000 cases worth nearly 1 billion GBP in the United Kingdom in the year of 2011. Even though insurance companies and related organizations take numerous measures to combat this wide-spread phenomenon, due to its secretive nature, a major part of fraud goes undetected, resulting in an estimated total of another 2 billion of excess payments each year in the United Kingdom (see Association of British Insurers (2012)).

In light of its prevalence and economic extent, several insurance companies established their own investigative units to uncover insurance fraud. Being equipped with limited budgets, however, they are forced to verify only those claims which exhibit a comparatively high probability of containing fraud and a relatively high saving potential rather than analyzing every single incoming claim. A recent survey conducted by Coalition Against Insurance Fraud (2012) among 74 mostly property/casualty insurers revealed that 88% of the respondents employ technologies to support their investigators, two of the most common being "automated red flags" and "scoring capabilities".

Nevertheless, many insurers are just beginning to discover the necessity of establishing fraud investigation divisions within their own company. Interviews with experts in this field have revealed that, in particular, smaller insurance companies may not deem it worthwhile to invest in costly software, still relying on their intuition when it comes to detecting fraudulent claims.

Our aim is, hence, to identify the determinants that would make it possible to draw conclusions on the likelihood of a claim seeking unfounded indemnification. Based on such a catalog of criteria, insurance companies would be able to use their limited resources to reveal defrauding attempts more effectively. In addition, honest policyholders may also benefit from an improved auditing scheme and hit ratio. Processing times would likely shorten, thereby resulting in reduced waiting periods for indemnification.

With the aim of detecting insurance fraud by engaging in auditing processes, the insurance companies' strategy can be assigned to the category of costly state verification (see, e.g., Townsend (1979), Mookherjee and Png (1989), Bond and Crocker (1997), Picard and Fagart (1999), Dionne, Giuliano, and Picard (2009)). The latter is based on the assumption of information regarding the (allegedly) insured event being distributed asymmetrically between the policyholder and the respective insurance company. It is, therefore, possible that the policyholder may misrepresent facts and figures in order to obtain a higher or even unjustified indemnification. To confront and discourage any defrauding attempts, insurance companies perform verification processes to determine the truthfulness of incoming claims and may then choose to impose penalties. Since audits, however, incur costs and the respective divisions have limited funds at their disposal, a choice must be made as to which of the incoming claims to test. An important consideration in this context is the weighing of incurred costs against potential savings related to detected fraud attempts.

An alternative approach in the handling of insurance fraud is subsumed under the term costly state falsification (see, e.g., Crocker and Morgan (1998), Crocker and Tennyson (2002)). Other than in the first-mentioned one, the idea behind this approach is for the policyholder to be able to manipulate a claim at monetary expense such that the fraud attempt becomes undetectable. In this case, auditing proves to be obsolete, leaving the insurer with the potential option of indemnifying all incoming claims without further verification while at the same time raising the premium payments. This approach would be in line with the findings of Clarke (1989) and Morley, Ball, and Ormerod (2006), who revealed that insurance companies were concerned with reputational risks as a consequence of excessive auditing. Such an approach could create a negative image in public perception, as a result of which individuals may be tempted to switch to one of the company's competitors.

For the purpose of our study, we will use the term "fraud" or "fraud attempt" as a collective term for all those cases within our data sample for which the respective insurance company has found sufficient evidence to categorize them as such. The phenomenon of insurance fraud having many facets, there is a variety of forms that may be observed in this context (see, e.g., Picard (2001), Crocker and Tennyson (2002), Tennyson (2008)).

Based on the severity of the offense, a common distinction is made between soft fraud and criminal/ hard fraud. According to Derrig (2002), criminal fraud is defined as the "willful act of obtaining money or value from an insurer under false pretenses or material misrepresentations". Expert interviews as well as previous research (see, e.g., Weisberg and Derrig (1991), Viaene and Dedene (2004), Tennyson (2008)), however, have revealed that the majority of defrauding attempts is situated in an ethical gray area rather than containing outright fraud. Even in the absence of a definition, the term soft fraud is related to attempts to inflate the claims amount after the occurrence of an insured event in order to obtain higher indemnification.

Aside from policyholders, other potential actors associated with the occurrence of insurance fraud include insurance brokers, intermediaries and service providers (see, e.g., Dulleck and Kerschbamer (2006) and International Association of Insurance Supervisors (2011)). Whether it is charging excessive prices or providing unnecessary services and treatments, such activities can be performed either with or without the knowledge of the respective policyholder aiming to obtain additional payments from the insurance company (see Tennyson (2008)).

The aim of our study is to identify the determinants that help to accurately distinguish between legitimate and illegitimate incoming claims. Hereby, we take into account characteristics regarding the policyholder himself, the insured vehicle, the signed policy and the loss event itself.

Previous literature has analyzed potential indicators that predict the likelihood of fraud by employing discrete choice models.[1] For the specifics of the US insurance market, Tennyson and Salsas-Forn (2002) as well as Derrig, Johnston, and Sprinkel (2006) analyze the phenomenon of insurance fraud related to automobile personal injuries requiring medical treatment. While the latter present some exemplary measures to handle fraud attempts, Tennyson and Salsas-Forn (2002) find that auditing processes contain both a

[1] For different approaches to determining fraud indicators see, e.g., Derrig, Weisberg, and Chen (1994) and Brockett, Derrig, Golden, Levine, and Alpert (2002). Ai, Brockett, Golden, and Guillén (2013) use such indicators to determine the overall fraud rate in a population of filed claims.

detection and a deterrence component. Furthermore, Belhadji, Dionne, and Tarkhani (2000) identify fraud indicators to determine their actual impact on the fraud probability of a claim using a representative data set from Canadian insurance companies. A slightly different path in this context is followed Dionne et al. (2009). Using the scoring approach, they derive a red flag strategy indicating which of the suspicious claims should be referred to an external investigative units. The result is an optimal auditing strategy in the face of a cost-minimizing insurance company.

Apart from that, Artís, Ayuso, and Guillén (1999), Artís, Ayuso, and Guillén (2002), Caudill, Ayuso, and Guillén (2005), Pinquet, Ayuso, and Guillén (2007) and Bermúdez, Pérez, Ayuso, Gómez, and Vázquez (2008) address potential issues which may surface in relation to the data sample itself. These include selection biases based on the insurers' own criteria for selecting claims to undergo auditing in the first place (see Pinquet et al. (2007)) and oversampling of fraudulent claims in the data set (see Artís et al. (1999) and Bermúdez et al. (2008)). Furthermore, Artís et al. (2002) and Caudill et al. (2005) account for misrepresentation of honest claims, i.e., cases that the insurance company mistakenly considers as legitimate.

With this paper, we aim to extend the existing studies on the identification of insurance fraud. Based on the literature, we develop a number of hypotheses to gain new insights into the drivers of fraudulent behavior. Furthermore, we utilize of a comprehensive data set from the automobile insurance market in Switzerland. To the best of our knowledge, such an analysis on indicators predicting the existence of insurance fraud has not been performed for the Swiss market to date.

The data sample we acquired for our analysis is comprised of audited claims from a major Swiss insurance company. The audits were performed throughout the time period between 2004 and 2011 within their automobile devision. Potential fraud indicators are available on the policyholder, vehicle, policy and loss level. By applying logistic regression methods, we determine which characteristics have a significant impact on the occurrence of fraud and could therefore be used to trigger auditing processes.

One particular interesting result refers to the impact of the insured loss amount on the policyholder's decision to engage in fraudulent activities. We are able to show that the option to defraud one's insurance company is solely taken into consideration for comparably small loss amounts, proving that behavioral adaption in the context of insurance fraud does take place.

Particularly from a practical perspective, the identification of factors revealing the probability of defrauding attempts is crucial. Being able to assess the fraud potential of an incoming claim is an essential step in the claims settlement process. Since the resources that are set aside to combat insurance fraud are limited, it is of great importance to distinguish between those claims for which verification is deemed sensible and those which should be paid out right away. This paper's derived catalog of criteria can serve as a basis for implementing auditing strategies to handle defrauding attempts more effectively. The extent to which insurance companies make use of this information, however, depends particularly on their available budgets.

The remainder of this paper is structured as follows: Section 2 sets forth ten hypotheses with regard to potential fraud indicators and their respective effect on the likelihood of committing fraud. We then provide a comprehensive overview of our data sample using descriptive measures, before presenting our theoretical model. The results of the logistic regressions are reported and discussed in Section 3. Finally, in Section 4, we summarize our findings and provide an outlook for future research.

2 Theory and Hypotheses Development

2.1 Development of Hypotheses

In the following, we develop several hypotheses with regard to the determinants that may serve as potential fraud indicators revealing the probability of an incoming claim being untruthful. Using data from auto insurance policies of a Swiss insurance company, we take into consideration characteristics on the policyholder, vehicle, policy and loss level.

Previous literature has already analyzed suitable indicators in the context of insurance claims fraud. In our study, we pick up the presented research results to examine whether or not the fraudulent claims in our data sample exhibit the same characteristics. In addition, several additional hypotheses are introduced which, to the best of our knowledge, have not yet been tested empirically.

Fraud Indicators Based on Policyholder Characteristics

Policyholder Age According to a representative population survey commissioned by the German Insurance Association GDV (2011), there is a wide-spread perception among all age groups that defrauding one's insurance company would generally be easy. A closer look, however, reveals this attitude to be slightly more prevalent among younger policyholders than older ones. Similarly, a study published by the Insurance Fraud Bureau (2012) reveals that while 8% of all survey participants stated their willingness to participate in a staged accident for financial profit, this number increases to 14% among young people. One reason behind this attitude may be that financial benefits from successful fraud attempts carry more weight for younger policyholders than for older ones due to their respective average assets. These elaborations are also in line with the findings of Artís et al. (2002) who show in their data sample that younger drivers are more likely to try to defraud their insurance company. Therefore, we hypothesize:

H_1: *The younger the policyholders are, the more likely they are to engage in fraudulent activities.*

Fraud Indicators Based on Vehicle Characteristics

Vehicle age In connection with characteristics related to the insured vehicle itself, its age may be of interest to predicting the probability of a claim being fraudulent. Artís et al. (2002) were able to prove this link, empirically showing that older vehicles are more likely to be involved in fraudulent activities since policyholders may perceive its cash value as a form of additional funds when purchasing a new car. Following this line of reasoning, one can assume fraud in this context not only to occur in the form of build-up, but also as seeking indemnification for uninsured events in order to gain financial benefits. We include this aspect in our study, and hypothesize:

H_2: *The older vehicles are, the more likely they are to be involved in insurance claims fraud.*

Vehicle type Additionally, the vehicle's class may be associated with a particular probability of being involved in fraudulent activities. In our data set, we can distinguish between regular passenger cars, transporters and motorcycles. Insurers have long been known to take the vehicle class into account when pricing the policy since it serves as an indicator for driving behavior and related accident frequency. Therefore, we include this variable in our analysis and postulate:

H_3: The class of an insured vehicle has a significant impact on the probability of filing a fraudulent claim.

Vehicle value Another characteristic related to the vehicle's characteristics is its value, which is composed of its catalog price and the value of any accessories, such as audio systems, car phones or air conditioning. In particular, these additions, whether fitted already by manufacturer or at some later point, have the potential to substantially increase the insured vehicle's value the consequence being higher insurance premiums. As policyholders then have financial incentives to engage in fraudulent activities, we aim to verify the following hypothesis within our data sample:

H_4: The higher the value of an insured vehicle, the more likely defrauding attempts become.

Leasing More and more individuals are choosing to lease their automobiles instead of purchasing them. A recent representative study in Switzerland commissioned by comparis.ch (2011), the leading Swiss Internet comparison service, revealed that the share of leased vehicles accounts for 14% of the overall private automobile market. This number rises even up to 23% with regard to the share of leased cars among all new private ones. With an average price of 42,328 CHF, leased cars are, on average, slightly more expensive than those paid for in cash costing 40,091 CHF. Leasing contracts usually provide the lessee with the right to purchase the then-used vehicle at the end of contract. Since the price is generally determined already by the time of signing the leasing agreement, it is in the lessee's interest to obtain the car in its best possible condition. This, however, may incentivize individuals to misuse their insurance coverage, to eliminate defects of any kind at the expense of the insurance company. As a consequence, one could expect the magnitude of claims to be disproportionally high for leased vehicles. Therefore, we hypothesize:

H_5: Leased vehicles are more likely to be engaged in fraudulent activities than purchased ones.

Fraud Indicators Based on Policy Characteristics

Loss-free An individual's perception of insurance in general may serve as an incentive in the context of fraud. Surveys have discovered that policyholders perceive build-up in particular as a way to obtain a compensation for former premium payments without having made a claim (see, e.g., International Association of Insurance Supervisors (2011), Miyazaki (2008), Duffield and Grabosky (2001)). This attitude adopts the common idea of treating insurance as an investment which has to eventually pay off. Consequently, we expect policyholders who have been in an insurance relationship for several periods without filing a claim to use the opportunity to inflate the amount of an insured loss by the time of its occurrence. We therefore postulate:

H_6: The longer the insurance relationship exists while remaining loss-free, the more likely defrauding attempts become.

Records In the context of automobile insurance, most insurance companies offer their policyholders bonus-malus policies providing them an incentive not to file claims for all kinds of minor losses and at the same time rewarding them for accident-free driving records (see, e.g., Moreno, Vázquez, and Watt (2006)). We believe that bonus-malus policies may be an obstacle in filing a claim to begin with, particularly for small damages. Since, however, it implies negative consequences for the policyholder in the form of increased premium payments for the consecutive period, this penalty may at the same time provide an incentive to obtain additional payments from a claim in order to compensate for additional future expenses. This kind of attitude is expected to be particularly observable among individuals having a bad driving record since they already are likely to be at the highest premium level. In these cases, Artís et al. (1999) argue that the claimants may feel like they have "nothing to lose" anyway. Based on their data sample, Artís et al. (1999) are able to show that the number of previous claims indeed has an impact on the likelihood of a fraud attempt. We therefore aim to verify the following:

H_7: The higher the number of previous claims, the higher the likelihood of a claim containing fraud.

Fraud Indicators Based on Loss Characteristics

Type of damage In discussions with experts from several fraud investigation divisions, attention was drawn to the different types of damages for which policyholders file claims. Particular focus was placed on loss events whose magnitude may easily be manipulated by either "overprovision" or "overcharging" (see Tennyson (2008)) as well as to damages that are allegedly difficult to verify and, hence may encourage defrauding attempts. These include, among others, glass breakage and collisions. Therefore, we include this variable in our analysis and postulate:

H_8: Types of damages which are deemed to be difficult to verify (e.g., glass breakage and collisions) are more likely to contain fraud than those which are deemed easily verifiable.

Loss amount In filing a fraudulent claim, its magnitude is of particular importance. We are convinced that policyholders do have a presentiment of the existence of auditing and hence take it into consideration when engaging in fraudulent activities. With claims of high magnitude being supposedly one of the targets under investigation, we expect fraud prone policyholders to contemplate such actions solely in cases of smaller-valued loss events and in the form of a percental surcharge on the actual loss amount. This approach would additionally leave the option to excuse any incorrect claims as a mistake if audited by the insurance company(see, e.g., Emerson (1992)). Hence, we hypothesize the following:

H_9: Smaller-valued claims are more likely to contain some kind of fraud than higher valued ones.

Delay Previous studies have shown (see, e.g., Artís et al. (2002), Dionne et al. (2009)) that the longer the lag between the accident and the filing of a report to the insurance company, the higher the likelihood of the respective claim containing some kind of fraud. The reason behind this observation is assumed to be that policyholders take this time to elaborate on the alleged story that they are trying to sell to their insurance company. We therefore include this aspect in our analysis, and postulate:

> H_{10}*: Greater delay in reporting an event to the insurance company increases the probability of fraud attempts being undertaken.*

2.2 Data Set

Our data set is constituted of personal-, vehicle-, policy- and loss-related information on the population in the automobile insurance division of a Swiss insurance company. It is comprised of all claims filed between the years of 2004 and 2011, summing up to a total of $1,429,896$ claims seeking for almost 2.5 bn CHF in indemnification. Throughout this time period, $7,407$ (0.52 percent) of those claims were examined by the company's fraud investigation division. The indemnification payments for these cases summed up to a total of more than 60 mn CHF. Among the $7,407$ audited claims, 402 (5.43 percent) were identified as fraudulent, mainly exhibiting signs of build-up. Consequently, the majority of these claims received some partial indemnification, only 1.49 percent of them were denied any payment.

As indicated previously, we make use of the word "fraud" as a collective term for all cases that were categorized as such by the insurance company. This, however, does not imply that every single one of these cases is an offense in the criminal-law sense. Judging from the high amount of partial indemnifications, the majority of audited claims seems to have exaggerated the actual loss amount rather than completely forging an insured loss event. Hence, these cases would fall into the category of soft and not criminal fraud. Nevertheless, in our study we choose not to make a distinction regarding the extent to which the individuals defrauded the insurance company.

Furthermore, we do not differentiate between the potential actors in the context of insurance fraud. However, besides the policyholders themselves, third parties like repair shops may also be involved in fraudulent activities. On the one hand, the initiative may be taken by the insured hoping for previous damages, unrelated to the current accident, to get repaired. On the other hand, the repair shops may be the ones to inflate the loss amount, either by charging overly high prices or providing unnecessary services (see Tennyson (2008)). These actions can be undertaken with or without the knowledge of the policyholder.

Data Selection

The insurance company's decision as to whether an incoming claiming has to undergo verification or not was based on personal evaluation of the incoming cases. The investigation division consists primarily of employees with a police background having broad experience with fraudulent activities in the insurance context. A predefined set of fraud indicators, however, that may serve as hints for the probability of fraud being present in a claim, had not existed during the time period between 2004 and 2011.

Nevertheless, it can be expected that the investigators did not proceed arbitrarily. Being aware of the limited resources at their disposal, they sought to focus on those claims that appeared to have a high probability of being illegitimate and that exhibited a high saving potential. Even in the absence of a predefined set of selection criteria, they likely chose the claims for auditing accordingly. These criteria, however, whether chosen deliberately or not, may influence the composition of our data sample of audited claims and therefore impact the results of the regression analyses.

For this purpose, we report measures on sample composition for the sample of all filed claims as well as the subsamples of audited and not audited claims. The results can be found in Tables 11 and 12 in Appendix B. According to the results, investigators seem to have selected disproportionally young policyholders who drive either older or high-valued vehicles. They exhibited flawless driving records, however by the time of loss occurrence, seeking comparably high indemnification. Regarding the type of damage, cases reporting the theft of the insured vehicle seem to have been the target of investigations.

Selection bias being probably present to some extent, we nevertheless perform our analyses to identify potential fraud indicators. Insurance companies being driven by the need to minimize their cost and time consumption, it seems unrealistic to expect anyone to perform auditing on a completely random basis. We will therefore not be able to acquire a data sample that is free of all selection biases.

2.3 Descriptive Statistics

In this section, we present descriptive measures to provide insight into the full data samples of audited claims as well as the subsamples of fraudulent and legitimate ones. Tables 1 and 2 give an overview of all variables used in our analysis.

	Audited Claims		**Defrauders**		**Non-defrauders**		
	mean	s.d.	mean	s.d.	mean	s.d.	p-value
	N=7407		N=402		N=7005		
Policyholder age	39.18	13.87	40.78	12.52	39.09	13.94	0.0111
Vehicle age	7.39	5.75	6.54	4.68	7.43	5.80	0.0004
Vehicle Value (CHF)	48,313	59,250	51,929	39,809	48,105	60,171	0.0707
No. consec. loss-free years	4.26	2.33	4.84	2.53	4.22	2.31	< 0.0001
No. previous records	3.18	7.55	2.11	1.84	3.24	7.75	< 0.0001
Loss amount (CHF)	8,711	16,996	5,379	12,088	8,847	17,153	< 0.0001
Delay in filing claim (days)	15.90	43.61	13.23	35.39	16.06	44.05	0.1462

Table 1: Descriptive Statistics for the Sample Composition

This table reports the mean and standard deviation (s.d.) of different characteristics related to policyholder, vehicle, policy and loss with regard to the full sample of audited claims. This information is narrowed down particularly for the two subsamples of proven fraud attempts (i.e., defrauders) and legitimate claims (i.e., non-defrauders). Furthermore, the last two columns provide the results of a two-sample t-test.

The first column in Table 1 shows the mean and standard deviation for a number of policyholder-, vehicle-, policy- as well as loss-related characteristics for the overall data set. Policyholders whose claims had to undergo verification were on average just over 39 years old. The vehicles involved in the loss events were a little over 7 years of age, being worth more than $48,000$ CHF including accessories. While the claimants had remained loss-free for over 4 consecutive years, their driving records were comprised of 3 previous loss events. On average, it took insured individuals almost 16 days to file a claim after the loss event occurred, seeking over $8,700$ CHF of indemnification.

The second and third columns in Table 1 specify this information for the subsamples of fraudulent and legitimate claims in order to uncover potential differences between these two groups. To identify whether any discrepancies are the result of significant differences between the subsamples or simply arise randomly, we perform a two-sample t-test for the equality of the means (see p-values). While policyholders proven to have engaged in fraudulent activities were on average over 40 years old, honest ones were nearly two years younger. With the corresponding p-value of the t-test being 0.0111, i.e., less than 0.05, this observation may be a hint that the policyholder's age serves as an indicator of the existence of fraud. Regarding their vehicles, it is striking that those participate in a defrauding attempt were approximately one year younger but almost 4000 CHF more expensive than those in the opposing group. Again, taking the results of the t-test into account, these variables might allow us to draw conclusions on the probability of fraud. Furthermore, in terms of driving behavior, there seem to be significant differences between the two subsamples. Claimants belonging to the group of defrauders have remained loss-free for almost one year longer and, at the same time, were involved in fewer accidents during the whole duration of their insurance relationship. Surprisingly, however, by the time of the loss occurrence, fraud-prone policyholders claimed loss events totaling to a little more than half the cost of that of their honest counter parts. Lastly, we observe that the delay in filing a claim appears to be irrelevant when predicting the probability of fraud.

Table 2 provides further information on the composition of the data set. Similarly to Table 1, we report the number and percentages of characteristics on the policyholder, vehicle, policy and loss level for the data set of audited claims in column one, and specify this information for the subsamples of fraudulent and legitimate claims in columns two and three respectively.

Comprising 58.90 percent of the whole population, Swiss citizens accounted for the majority of all policyholders. This number drops slightly (to 44.53 percent) among the subsample of fraud-prone claimants. While the greater portion of the overall policyholder population (72.09 percent) had their place of residence in the German-speaking part of Switzerland, only 22.82 percent indicated that their place of residence was among the French-speaking cantons of Switzerland and merely 5.09 percent in the Italian-speaking part.[2] These numbers do not seem to change considerably when comparing the subsamples of fraudulent and honest individuals. With respect to vehicle-related characteristics, we report the vehicle type and whether the latter was leased or not. A majority of about 71 percent of all policyholders had insurance coverage for a regular passenger car. This number rises by almost seven percentage points among the subsample of defrauding claimants. The opposite holds true for motorcyclists among the population. While their share among all

[2]Based on the prevalent quantity, the Swiss cantons are allocated as follows: Ticino to the Italian-speaking part, Geneva, Vaud, Neuchatel, Jura and Fribourg to the French-speaking part and the remaining cantons to the German-speaking part.

	Audited Claims		Defrauders		Non-defrauders	
	No.	Percent	No.	Percent	No.	Percent
Policyholder related characteristics						
CITIZENSHIP						
Swiss	4363	58.90	179	44.53	4184	59.73
other	3044	41.10	223	55.47	2821	40.27
Total	7407	100.00	402	100.00	7005	100.00
AREA OF RESIDENCE						
German-speaking part	5313	72.09	282	70.15	5031	72.20
French-speaking part	1682	22.82	83	20.65	1599	22.95
Italian-speaking part	375	5.09	37	9.20	338	4.85
Total	7370	100.00	402	100.00	6968	100.00
Vehicle related characteristics						
VEHICLE TYPE						
Car	3525	71.34	210	78.36	3315	70.94
Transport	262	5.30	15	5.60	247	5.29
Motorcycle	1145	23.36	43	16.04	1111	23.77
Total	4941	100.00	268	100.00	4673	100.00
LEASING						
Leased	1404	18.96	123	30.60	1281	18.29
Not leased	6003	81.04	279	69.40	5724	81.71
Total	7407	100.00	402	100.00	7005	100.00
Policy related characteristics						
BONUS PROTECTION CLAUSE						
Included	2991	40.38	203	50.50	2788	39.80
Not included	4416	59.62	199	49.50	4217	60.20
Total	7407	100.00	402	100.00	7005	100.00
Loss related characteristics						
TYPE OF DAMAGE						
Theft	2437	32.90	100	24.86	2337	33.36
Glass	1130	15.25	25	6.22	1105	15.77
Collision	1368	18.71	124	30.85	1244	17.76
Others	2472	33.37	153	38.06	2319	33.10
Total	7407	100.00	402	100.00	7005	100.00

Table 2: Descriptive Statistics for the Sample Composition

This table describes the sample composition using different categorical variables on the policyholder, vehicle, policy and loss level. Besides providing an overview of the complete data sample of audited claims, the information is further differentiated with regard to fraudulent (i.e., defrauders) and legitimate (i.e., non-defrauders) claims.

claimants sums to 23.36 percent, they only account for 16.04 percent of all detected fraud attempts. Transporters form the smallest part of all vehicle types comprising around 5.30 percent of the overall data sample as well as within the two subsamples. While less than 19 percent of all insured vehicles were leased, they account for about 30 percent of all cases proven to have engaged in fraudulent activities. Additionally, we provide information as to whether the policyholders had included a bonus protection clause in their contracts or not. This holds true for about 40 percent of the whole population, and over 50 percent among the subsample of detected defrauders. Finally, with respect to the claimed loss type, we distinguish between theft of the vehicle, glass breakage, collision and other damages.[3] Almost one third (32.90 percent) of all audited claims had reported the theft of the insured vehicle, whereas glass breakage and collision accounted for approximately 15 and 19 percent of all incidents, respectively. The shares of theft and glass breakage, however, drop notably by eight percentage points each within the subsample of fraudulent claims, while the portion of cases including collisions rises by more than twelve percentage points.

2.4 Model Derivation

The aim of our study is to identify the impact a set of explanatory (predictor) variables has on a dichotomous (binary) dependent variable, i.e., taking on solely one of the two values - one and zero (fraud and no fraud). We are hence envisaging the possibility of employing the logistic regression model.

In this relation, let us consider the following linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \epsilon_i = \mathbf{x_i}\boldsymbol{\beta} + \epsilon_i \tag{1}$$

where y_i denotes the outcome of the dependent variable for the i^{th} claim, i.e., fraud or no fraud, and x_{im} represents the value of the m^{th} explanatory variable for the i^{th} claim. Furthermore, β_m specifies the regression coefficients to be estimated, with β_0 being the intercept, and the random variables ϵ_i indicate the error terms. Having a system of linear equations, one may also abbreviate by using matrix notation. Hereby, $\mathbf{x_i}$ is a column vector with each row containing the values of the explanatory variables for the i^{th} claim, and $\boldsymbol{\beta}$ a column vector with the corresponding regression coefficients.

In contrast to linear regression models, however, the logistic regression does not pursue the estimation of the dependent variable's outcome y_i itself, but rather its probability of occurrence π_i which is defined as $\pi_i = \text{Prob}(y_i = 1) = \mathbb{E}(y_i)$ since y_i is dichotomous, i.e., Bernoulli distributed. Applying this to Equation (1), we obtain

$$\pi_i = \mathbf{x_i}\boldsymbol{\beta}, \tag{2}$$

since $\mathbb{E}(\epsilon_i) = 0$ for all i. This equation is commonly referred to as linear probability model.

Unfortunately, applying the ordinary least squares (OLS) method for estimating the regression coefficients

[3] The sub-category "others" comprises, among others, damages caused by hail, martens and other wild life, parking damages and theft of valuable left in the vehicle.

β_i, as usually done in relation to linear regression models, would give rise to a series of problems in this context. It may predict values outside the permitted range of $[0, 1]$, and is not able to capture heteroscedasticity and non-normality of error terms arising with dichotomous dependent variables (see, e.g., Pohlmann and Leitner (2003)). In addition to all this, utilizing OLS may produce nonsensical predictions for the estimation results. These obstacles can be overcome by drawing on the logistic model, which makes use of the logistic function $\text{Prob}(z) = 1/(1 + \exp(-z))$. With Equation 2, this results in

$$\pi_i = \text{Prob}(\mathbf{x_i}\boldsymbol{\beta}) = \frac{1}{1 + \exp(-\mathbf{x_i}\boldsymbol{\beta})}. \tag{3}$$

In this case, the regression coefficients β_i, also referred to as logit coefficients, are derived by means of maximum likelihood estimation (MLE). Its aim is to determine those parameter values β_i, which make the observation of the collected data y_i and $\mathbf{x_i}$ the most likely.

3 Empirical Results

In this section, we present and discuss the results of the logistic regression model introduced previously. Hereby, we begin with the set of explanatory variables with regard to policyholder characteristics and extend this initial model stepwise by adding the variables on the vehicle, policy and loss level in each iteration, respectively.

In order to compare the different models against each another and ultimately to assess how well the final model actually fits the observations, we present different measures of model adequacy.

3.1 Logistic Regression Results

Model 1: Policyholder Characteristics

The first model considers only the influence that the policyholder characteristics introduced in Tables 1 and 2 have on the decision to defraud the insurance company or not. Results of the logistic regression are reported in Table 3.

$N = 7002$	β_i	$\exp(\beta_i)$	s.e.	p-value	sig.
Constant	–3.5406	0.0290	0.1798	< 0.0001	***
Policyholder age	0.0098	1.0098	0.0038	0.0104	*
citizenship:other	0.6718	1.9578	0.1067	< 0.0001	***
area of residence:fr	–0.1945	0.8232	0.1338	0.1461	
area of residence:it	0.4960	1.6421	0.1926	0.0100	*

Table 3: Logistic Regression with Determinants Related to Policyholder Characteristics (Model 1)

Results for the logistic regression of the dependent variable (fraud/no fraud) with three explanatory variables on the policyholder level (plus constant). The regression coefficients β_i indicate the contribution of each explanatory variable on the logit, $\exp(\beta_i)$ the corresponding effect on the odds ratio, and s.e. represents the standard error of the respective determinant. Significance levels (sig.): *** = 0.1 percent, ** = 1 percent, * = 5 percent, . = 10 percent.

As already indicated by the result of the two-sample t-test in Table 1, the claimant's age has a significant effect on the likelihood of engaging in fraudulent activities ($p < .05$). The respective regression coefficient being positive, older policyholders are more fraud prone than younger ones ($\beta = .01$). This confirms our assumption that the policyholder age is indeed an indicator for detecting fraud, however, it contradicts H_1.

Furthermore, we find both citizenship and area of residence to be statistically significant. In particular, claimants not having the Swiss citizenship appear to be more involved in dishonest activities than their counter-group ($\beta = .67$, $p < .0001$). With regard to the categorical variable area of residence, we use all policyholders living in the German-speaking part of Switzerland as a reference group. While individuals from the French-speaking cantons do not exhibit significantly different defrauding behavior ($\beta = -.19$, $p > .1$) compared to the German-speaking ones, claimants from the Italian-speaking part have a higher probability of engaging in fraudulent activities ($\beta = .50$, $p < .05$).

Model 2: Policyholder and Vehicle Characteristics

In addition to the policyholder characteristics, the second model takes the variables on the vehicle level into account (see Tables 1 and 2). The results of the corresponding logistic regression can be found in Table 4.

$N = 5156$	β_i	$\exp(\beta_i)$	s.e.	p-value	sig.
Constant	–2.772	0.0625	0.2694	< 0.0001	***
policyholder age	0.0059	1.0059	0.0047	0.2060	
citizenship:other	0.5955	1.8139	0.1255	< 0.0001	***
area of residence:fr	–0.1285	0.8794	0.1580	0.4159	
area of residence:it	0.3551	1.4263	0.2204	0.1071	
vehicle age	–0.0296	0.9708	0.0201	0.1424	
vehicle type:transporter	–0.0852	0.9183	0.3231	0.7920	
vehicle type:motorcycle	–0.3243	0.7230	0.0208	0.1196	
Vehicle value	0.0001	1.0000	0.0001	0.9459	
leasing:no	–0.4962	0.6088	0.1512	0.0010	**

Table 4: Logistic Regression with Determinants Related to Policyholder and Vehicle Characteristics (Model 2)

Results for the logistic regression of the dependent variable (fraud/no fraud) with four explanatory variables on the vehicle level in addition to the three policyholder characteristics (plus constant). The regression coefficients β_i indicate the contribution of each explanatory variable on the logit, $\exp(\beta_i)$ the corresponding effect on the odds ratio, and s.e. represents the standard error of the respective determinant. Significance levels (sig.): *** = 0.1 percent, ** = 1 percent, * = 5 percent, . = 10 percent.

Examining the regression coefficients and p-values for the policyholder characteristics, we find that result from Model 1 regarding the impact of citizenship is confirmed.

Furthermore, we see that, in this model set up, the vehicle's age may not serve as an indicator for fraud being prevalent in a claim ($\beta = -.03$. $p > .1$). With regard to vehicle type, the reference group is comprised of all cases where a regular passenger car was involved in the loss event. In comparison, neither transporters ($\beta = -.09$, $p > .1$) nor motorcyclists ($\beta = -.32$, $p > .1$) exhibited a significantly different defrauding behavior than drivers of passenger cars. The criterion of whether the vehicle was leased or not seems to be a good indicator of fraud ($\beta = -.50$, $p < .005$). With the regression coefficient being negative, we can conclude that owners of non-leased vehicles are less tempted to defraud the insurance company. This finding supports the assumption stated in hypothesis H_5. Surprisingly, however, the vehicle's value does not appear to be statistically significant for the existence of fraud ($p > .1$).

To evaluate whether extending the initial Model 1 by the variables on the vehicle level improves the fit, we conduct a likelihood ratio test. The results are reported in Table 7. According to the corresponding values ($\chi^2 = 24.7$, $p < .0001$), Model 2's fit proves to be significantly better.

Model 3: Policyholder, Vehicle and Policy Characteristics

The next step in extending our model is to additionally include all variables on the policy level. Table 5 shows the corresponding results.

$N = 5156$	β_i	$\exp(\beta_i)$	s.e.	p-value	sig.
Constant	–2.6110	0.0735	0.3580	< 0.0001	***
policyholder age	0.0063	1.0063	0.0048	0.1859	
citizenship:other	0.5163	1.6758	0.1272	< 0.0001	***
area of residence:fr	–0.1862	0.8301	0.1584	0.2399	
area of residence:it	0.3675	1.4441	0.2243	0.1014	
vehicle age	–0.0151	0.9850	0.0184	0.4097	
vehicle type:transporter	–0.0529	0.9485	0.3254	0.8709	
vehicle type:motorcycle	–0.5520	0.5758	0.2117	0.0091	**
vehicle value	0.0001	1.0000	0.0001	0.6818	
leasing:no	–0.5370	0.5845	0.1505	0.0003	***
No. consecutive loss-free years	0.0664	1.0687	0.0280	0.0178	*
No. previous records	–0.2506	0.7783	0.0422	< 0.0001	***
bonus protection clause	0.3362	1.3996	0.1360	0.0134	*

Table 5: Logistic Regression with Determinants Related to Policyholder, Vehicle and Policy Characteristics (Model 3)

Results for the logistic regression of the dependent variable (fraud/no fraud) with three explanatory variables on the policy level in addition to the three policyholder and four vehicle characteristics (plus constant). The regression coefficients β_i indicate the contribution of each explanatory variable on the logit, $\exp(\beta_i)$ the corresponding effect on the odds ratio, and s.e. represents the standard error of the respective determinant. Significance levels (sig.): *** = 0.1 percent, ** = 1 percent, * = 5 percent, . = 10 percent.

With regard to the parameters already included in Models 1 and 2, we see that our previous results are confirmed. The explanatory variables on the policyholder and vehicle level remain significant with respect to the prediction of the likelihood of fraud.

Both the number of consecutive loss-free years and the number of previous records seem to be statistically significant in predicting the probability of fraud. In particular, claimants who remained loss-free for a longer time period are more likely to get involved in fraudulent activities once they experience an insured loss event ($\beta = .07$, $p < .01$). This result provides proof for hypothesis H_6. Regarding driving records, we find that the less previous claims a policyholder has had the higher the likelihood of defrauding the insurance company ($\beta = -.25$, $p < .0001$). On the one hand side, this observation is consistent with the result from the number of consecutive loss-free years. On the other hand side, it seems counter-intuitive since a high driving record is associated with individuals having a bad and/or insecure driving behavior and therefore is deemed to be a predictor for the likelihood of fraud. Another indicator for the presence of fraud in a claim is the information whether a bonus protection clause was included in the insurance contract or not. According to the results of the logistic regression, policyholders who included this option in their contracts were involved in fraudulent activities more often than their counter-group ($\beta = .34$, $p < .05$).

The results of the likelihood ratio test of Model 3 against Model 2, presented in Table 7, confirm that the addition of the variables on the policy level does help to significantly increase the predictive accuracy ($\chi^2 = 29.10$, $p < .0001$).

Model 4: Policyholder, Vehicle, Policy and Loss Characteristics

Our final Model 4 reflects the effect of all variables on the predictability of fraud being present in a claim. The results are presented in Table 6.

$N = 4863$	β_i	$\exp(\beta_i)$	s.e.	p-value	sig.
Constant	–2.2030	0.1105	0.4343	< 0.0001	***
Policyholder age	0.0097	1.0097	0.0054	0.0793	.
citizenship:other	0.4635	1.5896	0.1519	0.0023	**
area of residence:fr	0.1347	1.1442	0.1879	0.4734	
area of residence:it	0.6636	1.9418	0.2555	0.0094	**
vehicle age	–0.0214	0.9788	0.0233	0.3585	
vehicle type:transporter	0.2390	1.2700	0.3505	0.4954	
vehicle type: motorcycle	–0.6660	0.5138	0.2768	0.0161	*
vehicle value	0.0001	1.0000	0.0001	0.0006	***
leasing:no	–0.7682	0.4638	0.1774	0.< 0.0001	***
no. consecutive loss-free years	–0.0011	0.9989	0.0337	0.9728	
no. previous records	–0.2206	0.8020	0.0460	< 0.0001	***
bonus protection clause	0.0065	1.0065	0.1661	0.9688	
type of damage:glas	–2.3110	0.0992	0.5363	< 0.0001	***
type of damage:collision	0.4653	1.5925	0.2096	0.0264	*
type of damage:other	–0.1971	0.8211	0.2125	0.3535	
loss amount	–0.0001	0.9999	0.0001	< 0.0001	***
delay in filing claim	–0.0077	0.9923	0.0035	0.0284	*

Table 6: Logistic Regression with Determinants Related to Policyholder, Vehicle, Policy and Loss Characteristics (Model 4)

Results for the logistic regression of the dependent variable (fraud/no fraud) with three explanatory variables on the loss level in addition to the three policyholder, four vehicle and three policy characteristics (plus constant). The regression coefficients β_i indicate the contribution of each explanatory variable on the logit, $\exp(\beta_i)$ the corresponding effect on the odds ratio, and s.e. represents the standard error of the respective determinant. Significance levels (sig.): *** = 0.1 percent, ** = 1 percent, * = 5 percent, . = 10 percent.

Starting with the loss characteristics, we find the magnitude of loss amount to be highly significant for filing fraudulent claims ($p < .0001$). The regression coefficient being negative implies that the smaller the loss amount, the more likely the existence of fraud. This observation was already indicated by the results from the two-sample t-test in Table 1 and provides proof for our predication stated in hypothesis H_9. The outcome for the delay in filing a claim, however, seems surprising. Even though its effect on the dependent variable is significant ($p < .05$), the negative sign of $\beta = -.01$ suggests that the shorter the time lag between the occurrence of loss and the report to the insurance company, the higher the likelihood for fraud. This is directly contrary to our assumption expressed in H_{10}. A possible explanation may be that, like in the case with the loss amount, claimants do suspect the insurance company to control for the delay in filing a claim and hence not only take the magnitude of loss into consideration when defrauding, but also manipulate the date of loss occurrence whenever it is possible. Once more, the aim is to feign realistic scenarios in order to not get audited and moreover detected.

On the policyholder level, the final model confirms some of the effects already predicted in Model 1: Older claimants have a higher probability of cheating on their insurance company. This observation, however, contradicts our assumption in hypothesis H_1. Put in the context of driving behavior and premium payments to date, a potential explanation may be that it is actually the older and thus (usually) more experienced policyholders who remain loss-free throughout long periods of time. Having paid insurance premiums over the course of many years, they may consider themselves long-standing customers who expect good will in form of generous indemnification in trade for their loyalty.

Regarding the variables related to the vehicle itself, the fully extended model once again confirms previous results: Both the vehicle value and the information whether the vehicle is leased or not are very good indicators for fraudulent claims. The proven effects support our assumptions expressed in hypotheses H_4 and H_5 respectively. Furthermore, we are able to show that the vehicle class has an impact on the likelihood of fraud, motorcyclists cheating less on the insurance company than drivers of regular passenger cars (see H_3). Solely, hypothesis H_2 does not hold true. Against our prediction, the age of the insured vehicle does not seem to have an impact on the likelihood to engage in fraudulent activities ($\beta = -.02$, $p > .1$).

Also, from the policy perspective, the results from Model 3 prove to be true: Exhibiting a small number of previous claims increases the likelihood of inflating the loss magnitude once an insured event occurs.

In comparison with Model 3, the fully extended Model 4 leads to a significantly better fit according to the likelihood ratio test in Table 7 ($\chi^2 = 106.55$, $p < .0001$), indicating that the fully extended Model 4 is to be preferred over the less evolved ones.

	$\Delta\chi^2$	Δdf	χ^2	df	p-value	sig.
Model 1	1651.1	4858				
Model 2	1626.4	4854	24.69	4	< 0.0001	***
Model 3	1597.3	4850	29.10	4	< 0.0001	***
Model 4	1490.8	4845	106.55	5	< 0.0001	***

Table 7: Likelihood Ratio Tests of the Models

Results of the pairwise likelihood ratio tests between the consecutive Models 1 to 4. Each extension of the previous model leads to a significant improvement in fit as indicated by the increasing values for χ^2 and the corresponding p-values. Significance levels (sig.): *** = 0.1 percent, ** = 1 percent, * = 5 percent, . = 10 percent.

Having decided on the final model for identifying fraud indicators, i.e., Model 4, we conclude this section by assessing its adequacy.

To check for potential problems with regard to the multicollinearity of independent variables within our data set, we determine their variance inflation factors, which can be found in greater detail in Table 10 in the Appendix. The fact that their values do not exceed the critical threshold value, the highest being 1.59 for the variable vehicle type, indicates that it is reasonable to assume our explanatory variables to be uncorrelated.

Table 8 displays the classification table for the full logistic regression model using all explanatory variables in our data set. The results confirm the good predictive accuracy of our model. In particular, with regard to the number of fraudulent claims, we are able to predict 73.76 percent correctly.This number, however, decreases slightly to 67.97 percent with respect to predicting the cases of legitimate claims. These figures indicate that our model is slightly more suitable for detecting fraudulent claims than honest ones. Apparently, many of the characteristics that help to identify fraudulent claims are also present among legitimate claims. On the one hand, the small number of detected fraudulent claims may be to blame. Only 402 out of the 7,407 audited claims (5.43 percent) were classified as fraudulent, making these cases rare events. On the other hand, our data sample was restricted to those criteria which are solely assumed to be helpful with respect to fraud detection. In light of this, there may exist other explanatory variables besides those included in our data sample, which may improve the distinction between honest and dishonest claims. These may include many determinants already known to the insurance company, but also some that have not been gathered yet.

	Predicted		
Observed	Fraud	No Fraud	% Correct
Fraud	149	53	73.76
No Fraud	1493	3168	67.97

Table 8: Classification Table for Full Model

This classification table illustrates the predictive accuracy of the logistic regression model in Table 6 by showing how many of the observed values for the dependent variable (fraud/ no fraud) are correctly predicted. The full model correctly predicts 73.76 percent of the fraud attempts and 67.97 percent of the legitimate claims.

3.2 Special Focus on Loss Amount

As already revealed in the course of this section, fraud-prone policyholders do take the loss amount into consideration when deciding whether to actually engage in fraudulent activities or not. More precisely, the results of the logistic regression model as presented in Table 6 indicate that the two are inversely proportional to each other, i.e., actions to obtain higher indemnification tend to be undertaken when the insured loss amount is comparably small.

In this subsection, we focus on this particular fraud indicator and discuss its effect on the decision to cheat one's insurance company. For this purpose, Figure 1 illustrates the link between loss amount and the likelihood of fraud being present in an incoming claim. It must be noted that, for the purpose of this analysis, we consider solely the loss amount as a factor for predicting the existence of fraud. While taking other significant exploratory variables into account would certainly increase the overall accuracy in detecting defrauding attempts (see Table 8), it would not impact the link between the loss amount's magnitude and the likelihood of fraud.

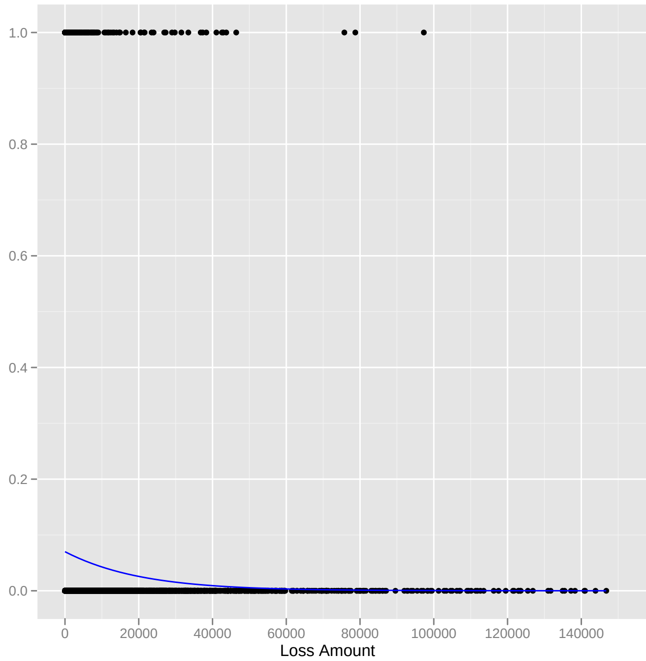

Figure 1: Contribution of Loss Amount to Predicting the Likelihood of Fraud. This figure illustrates both the actual magnitudes of loss events for honest and dishonest claims within our data sample as well as the loss amount's overall effect on the likelihood of fraud. While the blue line indicates the predicted probability of fraud depending on the magnitude of loss, each of the black points represents the loss amount of an actual claim from the data set, the corresponding y-value hinting fraud (1) or no fraud (0).

In Figure 1, we include the audited cases from our data set exhibiting a loss amount up to 150000 CHF. Each of the points in the figure depicts one claim within our data sample in terms of the corresponding loss amount and the information as to whether fraud was detected or not. Secondly, we add the predicted effect of the magnitude of a loss event on the likelihood to commit fraud, represented by the blue line in Figure 1.

This demonstates that, with regard to the honest claims, the loss amounts vary greatly across a wide range of magnitudes, reaching values of 150000 CHF and higher. For cases proven to be illegitimate, however, the contrary holds true. Here, the loss amounts seem concentrated in the range of up to 20000 CHF, occasionally going up as far as 40000 CHF. This observation is reflected by the predicted effect on the likelihood of fraud being involved in a claim. The blue line indicates that, while for small magnitudes,the loss amount accounts for almost ten percent of the predictability of fraud being present, this value drops rapidly to zero percent for loss amounts higher than 50000 CHF.

This leads us to conclude that defrauding attempts are not considered an option if the insured loss amount exceeds some threshold value. Primarily for relatively small magnitudes, some individuals may try to obtain higher indemnification payments from their insurance companies.

This observation provides proof for behavioral adaptation in the context of insurance fraud. Fraud-prone policyholders can be expected to adjust their defrauding schemes in such a way that their attempts remain undetected. Previous research has already shown that individuals fear negative consequences in the form of losses as a result of their actions more than they would appreciate a gain of the same size (see, e.g., Kahneman and Tversky (1979), Kerr (2012)). Hence, the overall objective is deemed to be perceived as presenting a legitimate claim rather than solely "maximizing" indemnity payments. Additional evidence for this link can be found, e.g., in Viaene, Derrig, Baesens, and Dedene (2002) and Tennyson (2008) who state that while only very few claims contain outright fraud, the majority of defrauding attempts is detected in cases seeking low to medium amounts of indemnification.

4 Conclusion and Critical Discussion

With fraud being identified as one of the central challenges in the industry to date and in the future, many insurance companies have established their own investigation divisions in the recent years. Nevertheless, many still rely largely on intuition when it comes to detecting wrongful claims. In our study, we identify criteria that allow for an accurate distinction between fraud-prone and honest policyholders and, by this means, predict the existence of fraud in a filed claim. Such a catalog of variables allows for a systematic approach to the combat against fraud, hopefully resulting in a higher detection rate. Moreover, it enables a targeted utilization of the limited resources that investigation divisions have at their disposal.

Our analysis is based on a sample of claims data comprised of 7,407 audited loss events in an automobile insurance division. The data was collected from a large Swiss insurance company between 2004 and 2011. The target variable being dichotomous, we employ logistic regression models to determine significant predictors of the presence of fraud in claims. The fit and adequacy of the different models, and particularly the final one, are assessed with the help of different measures. The analysis is rounded off with an in-depth examination of the effect of the loss amount on the likelihood of engaging in fraudulent activities.

The results of our logistic regression analyses portray fraud-prone policyholders as middle-aged individuals who prefer to drive high-valued cars and having signed a leasing contract more often than their honest counterparts. With regard to their driving behavior, individuals engaging in fraudulent activities prove to be rather experienced and safe drivers. They are characterized by having a low number of claims throughout their entire insurance relationship. Particularly noteworthy is the fact that they tend to file fraudulent claims for comparably small loss amounts, probably with the intention of remaining undetected. In a similar manner, they try to not attract attention by filing claims too late after the insured loss event occurred.

Another central result of this study is the empirical documentation of the link between loss amount and the probability of resulting fraud. As previous research has already suggested, the magnitude of an insured loss event has an inverse effect on the likelihood of taking fraudulent activities into consideration. The main reason for this observation can be policyholders' anticipation of auditing strategies to remain undetected and collect on the higher indemnification.

Our findings could be highly relevant to fraud investigators and underwriters alike. The information regarding fraud indicators can be utilized to perform auditing more effectively. Investigators would be given the opportunity to focus specifically on those claims which are deemed to have a high likelihood of being dishonest. Furthermore, the knowledge that some individuals are more prone to inflating loss amounts than others may be useful for other aspects of risk management as well. With regard to the pricing of insurance policies, the information on whether an individual should be counted among the fraud-prone or honest policyholders may be a relevant differentiation criterion. As part of the risk selection process, insurance companies may even decide not to provide coverage for individuals who can be expected to defraud the company to a large extent or who are unwilling to pay the corresponding insurance premium. These options are particularly interesting, since limited resources will prevent investigation units from verifying all incoming claims, even those which exhibit sufficient signs of the presence of fraud.

As with all studies, the current study has some limitations which may establish a basis for future research. Relying on the insurance company's decision as to which incoming claims to audit and which to indemnify right away, may have biased our view on potential fraud indicators. On the one hand, determinants which are identified based on a preselected sample predict the overall likelihood of fraud correctly. However, if they were already among the company's selection criteria, they may tend to overestimate the actual probability of its occurrence being amenable to the self-fulfilling prophecy. On the other hand, by disregarding part of the dishonest cases due to omission error, we may have been unable to capture additional fraud indicators, possibly even more suitable ones for predicting the existence of fraud in a claim.

5 Appendix A

Variable	**Definition**
Policyholder Characteristics	
Policyholder Age	Age of policyholder by the time of loss occurrence
Citizenship	Policyholder's citizenship (equals Swiss or other)
area of residence	Policyholder's area of residence within Switzerland (equals ge, fr or it)
Vehicle Characteristics	
Vehicle age	Vehicle age at the time of loss occurrence
vehicle type	Type of vehicle (equals car, transport or motorcycle)
Vehicle value	Value of vehicle including accessories (in CHF)
Leasing	Vehicle is leased (equals 1, otherwise equals 0)
Policy Characteristics	
no. consecutive loss-free years	Number of consecutive years without filing a claim
no. previous records	Total number of claims filed to date
bonus protection clause	Policy includes a bonus protection clause (equals 1, otherwise equals 0)
Loss Characteristics	
Type of damage	Type of damage policyholder seeks indemnification for (equals theft, glas, collision or others)
loss amount	Estimated loss amount of the claim (in CHF)
delay in filing claim	Time lag between occurrence of loss and filing claim to insurance company in days

Table 9: Explanatory variables used in the model

An overview of all variables and their respective definitions contained in our data set. We distinguish between variables on the policyholder, vehicle, policy and loss level. Based on this information, we perform logistic regression to determine potential indicators for the presence of fraud in a claim.

	vif
Policyholder age	1.076
citizenship	1.058
area of residence	1.154
vehicle age	1.345
vehicle type	1.587
vehicle value	1.255
leasing	1.324
no. consecutive loss-free years	1.154
no. previous records	1.170
bonus protection clause	1.260
type of damage	1.513
loss amount	1.202
delay in filing claim	1.022

Table 10: Variance Inflation Factors for All Explanatory Variables in Our Analyses

The variance inflation factors are employed to detect potential multicollinearity with regard to the explanatory variables. All corresponding values remaining below the critical threshold value - with 1.587 for the variable vehicle type being the maximum - this can be ruled out.

6 Appendix B

	Filed Claims		**Audited Claims**		**Not Audited Claims**		
	mean	s.d.	mean	s.d.	mean	s.d.	p-value
	N=1,429,896		N=7407		N=1,422,489		
Policyholder age	44.89	14.99	39.18	13.87	44.92	14.99	< 0.0001
Vehicle age	6.21	5.12	7.39	5.75	6.21	5.12	< 0.0001
Vehicle Value (CHF)	44,948	38,076	48,313	59,250	44,931	37,934	< 0.0001
No. consec. loss-free years	3.45	2.16	4.26	2.33	3.44	2.15	< 0.0001
No. previous records	4.67	33.02	3.18	7.55	4.68	33.10	< 0.0001
Loss amount (CHF)	1,775	3,760	8,711	16,996	1,740	3,535	< 0.0001
Delay in filing claim (days)	19.17	40.23	15.90	43.61	19.19	40.21	< 0.0001

Table 11: Descriptive Statistics for the Sample Composition

This table reports the mean and standard deviation (s.d.) of different characteristics related to policyholder, vehicle, policy and loss with regard to the full sample of filed claims. This information is narrowed down particularly for the two subsamples of audited claims and not audited claims. Furthermore, the last two columns provide the results of a two-sample t-test.

	Filed Claims		Audited Claims		Not Audited Claims	
	No.	Percent	No.	Percent	No.	Percent
Policyholder related characteristics						
Citizenship						
Swiss	1,055,362	73.98	4,363	58.90	1,050,999	74.06
other	371,131	26.02	3,044	41.10	368,087	25.94
Total	1,426,493	100	7,407	100	1,419,086	100
Area of residence						
German-speaking part	1,001,286	70.38	5,313	72.09	995,973	70.37
French-speaking part	338,016	23.76	1,682	22.82	336,334	23.76
Italian-speaking part	83,340	5.86	375	5.09	82,965	5.86
Total	1,422,642	100	7370	100	1,415,272	100
Vehicle related characteristics						
Vehicle type						
Car	980,258	92.24	3,525	71.34	976,733	92.33
Transport	52,930	4.98	262	5.30	52,668	4.98
Motorcycle	29,588	2.78	1,154	23.36	28,434	2.69
Total	1,062,776	100	4,941	100	1,057,835	100
Leasing						
Leased	340,247	23.80	1,404	18.96	338,843	23.82
Not leased	1,089,649	76.20	6,003	81.04	1,083,646	76.18
Total	1,429,896	100.00	7,407	100.00	1,422,489	100.00
Policy related characteristics						
Bonus protection clause						
Included	746,440	52.20	2,991	40.38	743,449	52.26
Not included	683,456	47.80	4,416	59.62	679,040	47.74
Total	1,429,896	100.00	7,407	100.00	1,422,489	100.00
Loss related characteristics						
Type of damage						
Theft	9,049	0.63	2,437	32.90	6,612	0.47
Glass	367,832	25.77	1,130	15.26	366,702	25.83
Collision	328,867	23.04	1,368	18.47	327,499	23.07
Others	721,376	50.55	2,472	33.37	718,904	50.64
Total	1,427,124	100.00	7,407	100.00	1,419,717	100.00

Table 12: Descriptive Statistics for the Sample Composition

This table describes the sample composition using different categorical variables on the policyholder, vehicle, policy and loss level - for all filed claims and with regard to audited and not audited claims.

Bibliography

Ai, J., P. Brockett, L. Golden, and M. Guillén, 2013, A Robust Unsupervised Method for Fraud Rate Estimation, *Journal of Risk and Insurance*, 80(1):121–143.

Artís, M., M. Ayuso, and M. Guillén, 1999, Modelling different types of automobile insurance fraud behaviour in the Spanish market, *Insurance: Mathematics and Economics*, 24:67–81.

Artís, M., M. Ayuso, and M. Guillén, 2002, Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims, *Journal of Risk and Insurance*, 69(3):325–340.

Association of British Insurers, 2012, No Hiding Place: Insurance Fraud Exposed, Technical Report September.

Belhadji, B., G. Dionne, and F. Tarkhani, 2000, A Model for the Detection of Insurance Fraud, *Geneva Papers on Risk and Insurance - Issues and Practice*, 25(4):517–538.

Bermúdez, L., J. Pérez, M. Ayuso, E. Gómez, and F. Vázquez, 2008, A Bayesian Dichotomous Model with Asymmetric Link for Fraud in Insurance, *Insurance: Mathematics and Economics*, 42(2):779–786.

Bond, E. and K. Crocker, 1997, Hardball and the Soft Touch: The Economics of Optimal Insurance Contracts with Costly State Verification and Endogenous Monitoring Costs, *Journal of Public Economics*, 63(2):239–264.

Brockett, P., R. Derrig, L. Golden, A. Levine, and M. Alpert, 2002, Fraud Classification Using Principal Component Analysis of RIDITs, *Journal of Risk and Insurance*, 69(3):341–371.

Caudill, S., M. Ayuso, and M. Guillén, 2005, Fraud Detection Using a Multinomial Logit Model With Missing Information, *Journal of Risk and Insurance*, 72(4):539–550.

Clarke, M., 1989, Insurance Fraud, *British Journal of Criminology*, 29(1):1–20.

Coalition Against Insurance Fraud, 2012, The State of Insurance Fraud Technology: A study of Insurer Use , Strategies and Plans for Anti-Fraud Technology, Technical Report September.

Crocker, K. and J. Morgan, 1998, Is Honesty the Best Policy? Curtailing Insurance Fraud through Optimal Incentive Contracts, *Journal of Political Economy*, 106(2):355–375.

Crocker, K. and S. Tennyson, 2002, Insurance Fraud and Optimal Claims Settlement Strategies, *Journal of Law and Economics*, 45(2):469–507.

Derrig, R., 2002, Insurance Fraud, *Journal of Risk and Insurance*, 69(3):271–287.

Derrig, R., D. Johnston, and E. Sprinkel, 2006, Auto Insurance Fraud: Measurements and Efforts to Combat it, *Risk Management and Insurance Review*, 9(2):109–130.

Derrig, R., H. Weisberg, and X. Chen, 1994, Behavioral Factors and Lotteries Under No-Fault with a Monetary Threshold : A Study of Massachusetts Automobile Claims, *Journal of Risk and Insurance*, 61(2):245–275.

Dionne, G., F. Giuliano, and P. Picard, 2009, Optimal Auditing with Scoring: Theory and Application to Insurance Fraud, *Management Science*, 55(1):58–70.

Duffield, G. and P. Grabosky, 2001, The Psychology of Fraud, In *Trends and Issues in Crime and Criminal Justice*, 199. Australian Institute of Criminology.

Dulleck, U. and R. Kerschbamer, 2006, On Doctors, Mechanics, and Computer Specialists: The Economic of Credence Goods, *Journal of Economic Literature*, 44(1):5–42.

Emerson, R., 1992, Insurance Claims Fraud Problems and Remedies, *University Of Miami Law Review*, 46:907–973.

GDV, 2011, Versicherungsbetrug: aktuelle Entwicklungen, Muster und ihre Abwehr, Technical Report.

Insurance Fraud Bureau, 2012, Crash for Cash: Putting the brakes on fraud, Technical Report.

International Association of Insurance Supervisors, 2011, Application Paper on Deterring, Preventing, Detecting, Reporting and Remedying Fraud in Insurance, Technical Report.

Kahneman, D. and A. Tversky, 1979, Prospect Theory: An Analysis of Decision under Risk, *Econometrica*, 47(2):263–292.

Kerr, D., 2012, Exploring the Role of Pseudodeductibles in Auto Insurance Claims Reporting, *Journal of Insurance Issues*, 35(1):44–72.

Miyazaki, A. D., 2008, Perceived Ethicality of Insurance Claim Fraud: Do Higher Deductibles Lead to Lower Ethical Standards?, *Journal of Business Ethics*, 87(4):589–598.

Mookherjee, D. and I. Png, 1989, Optimal Auditing, Insurance, and Redistribution, *Quarterly Journal of Economics*, 104(2):399–415.

Moreno, I., F. Vázquez, and R. Watt, 2006, Can Bonus-Malus Alleviate Insurance Fraud?, *Journal of Risk and Insurance*, 73(1):123–151.

Morley, N., L. Ball, and T. Ormerod, 2006, How the detection of insurance fraud succeeds and fails, *Psychology, Crime \& Law*, 12(2):163–180.

Picard, P., 2001, Economic Analysis of Insurance Fraud, In *Handbook of Insurance*, 1997. Springer.

Picard, P. and M.-C. Fagart, 1999, Optimal Insurance Under Random Auditing, *Geneva Papers on Risk and Insurance Theory*, 24(1):29–54.

Pinquet, J., M. Ayuso, and M. Guillén, 2007, Selection Bias and Auditing Policies for Insurance Claims, *Journal of Risk and Insurance*, 74(2):425–440.

Pohlmann, J. and D. Leitner, 2003, A Comparison of Ordinary Least Squares and Logistic Regression 1, *OHIO Journal of Science*, 103(5):118–125.

Tennyson, S., 2008, Moral, Social, and Economic Dimensions of Insurance Claims Fraud, *Social Research*, 74(4):1181–1204.

Tennyson, S. and P. Salsas-Forn, 2002, Claims Auditing in Automobile Insurance: Fraud Detection and Deterrence Objectives, *Journal of Risk and Insurance*, 69(3):289–308.

Townsend, R., 1979, Optimal Contracts and Competitive Markets with Costly State Verification, *Journal of Economic Theory*, 21(2):265–293.

Viaene, S. and G. Dedene, 2004, Insurance Fraud: Issues and Challenges, *Geneva Papers on Risk and Insurance - Issues and Practice*, 29(2):313–333.

Viaene, S., R. Derrig, B. Baesens, and G. Dedene, 2002, A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection, *Journal of Risk and Insurance*, 69(3):373–421.

Weisberg, H. and R. Derrig, 1991, Fraud and Automobile Insurance: A Report on Bodily Injury Liability Claims in Massachusetts, *Journal of Insurance Regulation*, 9(4):497–541.