Received: 5 September 2020 | Revised: 23 March 2021 | Accepted: 27 March 2021

DOI: 10.1111/jori.12340

ORIGINAL ARTICLE

Journal of Risk and Insurance

# Near-miss telematics in motor insurance

**Montserrat Guillen**[1] | **Jens Perch Nielsen**[2] | **Ana M. Pérez-Marín**[1]

[1]Department of Econometrics, Riskcenter-IREA, Universitat de Barcelona, Barcelona, Spain

[2]Cass Business School, University of London, London, United Kingdom

**Correspondence**
Montserrat Guillen, Department of Econometrics, Riskcenter-IREA, Universitat de Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain.
Email: mguillen@ub.edu

**Funding information**
Ministerio de Ciencia e Innovación, Grant/Award Number: PID2019-105986GB-C21; Institució Catalana de Recerca i Estudis Avançats, Grant/Award Number: ICREA Academia; Fundación BBVA, Grant/Award Number: Research in Big Data

**Abstract**

We present a method to integrate telematics data in a pay-how-you-drive insurance pricing scheme that penalizes some near-miss events. We illustrate our method with a sample of drivers for whom information on near-miss events and claims frequency records are available. We discuss the implications for motor insurance ratemaking. Our pricing principle is to combine a baseline insurance premium with added extra charges for near-miss events indicating risky driving (or discounts) that can be updated on a weekly basis. This procedure provides an incentive for safe driving. In our real-case study illustration, hard-braking and acceleration events as well as smartphone use while driving increase the cost of insurance.

**KEYWORDS**

claims frequency, dynamic ratemaking, Poisson model, pricing

## 1 | INTRODUCTION

The amount of data available in the motor insurance industry is expanding enormously, with the emergence of numerous new sources of data via the internet and social networks. Indeed, new telematics devices provide vast amounts of information that create a challenge for traditional statistical techniques, or machine learning techniques (Barry & Charpentier, 2020). In this paper, we present a new methodology that breaks down telematics data into strategic components that can be used both for pricing and for attempting to alter driver behavior. By introducing the concept of "*near-miss telematics*," we change the focus of motor insurance from a low-frequency model based on claims to a high-frequency model based on near-misses and reduce the complexity of further data analytics for data scientists, statisticians, and actuaries. Near-miss telematics offers a solution to the big data problem of modern telematics information; it also provides a new way of presenting the car insurance bill, which now resembles a phone bill. This paper considers billing on a weekly basis, but there is no reason why bill information should not be processed in real-time, thus providing drivers with a direct and clear incentive to drive more safely to save insurance costs.[1]

To assess the association between near-misses and claims in a telematics motor insurance company, claim records for pay-how-you-drive (PHYD) policies must be consulted. Claims frequency in motor insurance is generally very low: For example, in many countries, it is below one claim at fault every 10 years. It is therefore difficult to develop dynamic insurance ratings based on observed claims only, while state-of-the-art telematics data are being collected. The problem is the slow dynamics of the claims data. In this paper, we show that a near-miss approach can transform dynamic insurance rating from a difficult problem based on low-frequency events to a straightforward exercise based on high-frequency events simply because near-misses are much more frequent than claims. A positive side effect of this new high-frequency rating system is that it is much more transparent to the policyholder and may act as an incentive to drive more safely. The car insurance bill will look more like a traditional phone bill, where one can follow the cost of each call that increases the flat rate fee; the insurance bill will specify the cost of every near-miss event caused by the driver that adds to the standard fee paid for insurance coverage. As an immediate example of a likely behavioral change, our approach will probably lead most drivers to stop any mobile use while driving.

In this article, we propose to use the past claims history available in an insurance company for a group of insured drivers with vehicles equipped with an on-board-device (OBD) to find evidence of the association between claim frequency records and currently observed near-misses. We use Poisson regression models to develop a basic pricing scheme based on the traditional claim frequency model and the near-miss data. This scheme will allow companies to calculate near-miss telematics rates by means of a basic price, which can be personalized with traditional rating factors (age, driving experience, zone, vehicle power, and so on), and most interestingly, a variable rate that includes penalizations for some dangerous near-miss events. In this way, drivers with higher frequencies of near-miss events pay more than the rest. This scheme is integrated in the claim frequency model to account for incoming telematics data, thus creating a natural dynamic pricing plan. We suggest that this ratemaking mechanism can be implemented on a weekly basis, but the price can also be accommodated at other time intervals (for instance, daily or monthly). In addition, summary information about near-miss behavior can be released to the customers to provide feedback on their driving style and ideally to improve their future behavior behind the wheel. A general structure of near-miss telematics ratemaking in motor insurance is provided in Figure 1.

We illustrate our methodology with an example taken from a real case study, where we found a positive association between past claim records and some observed weekly frequencies of near-misses. Based on these results, a new ratemaking mechanism applying claims frequency records and near-miss observations is proposed.

This paper is organized as follows. In Section 2, we present related literature. Section 3 describes telematics ratemaking methods and the billing process. In Section 4 we present the

[1]Gaming and the awarding of prizes for the best customers might be a field for telematics insurers to explore. Nevertheless, insurers should be aware of the possibility that drivers could behave less save due to certain incentives.

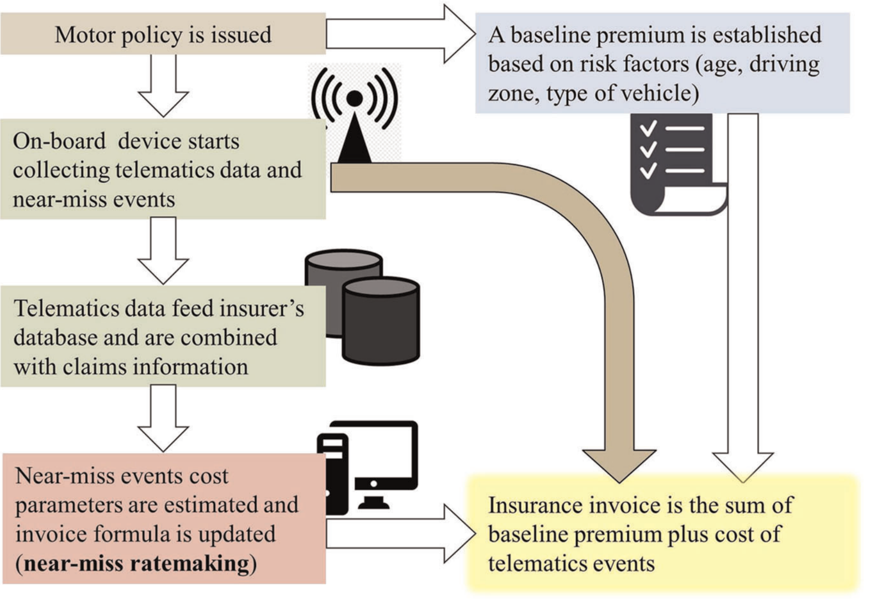


**FIGURE 1** Near-miss telematics ratemaking in motor insurance [Color figure can be viewed at wileyonlinelibrary.com]

real data set used in this article, show the model's results, and propose a pricing method based on telematics information. Finally, thśe last section concludes.

## 2 | NEAR-MISS TELEMATICS AND USAGE-BASED INSURANCE

A near-miss—a term borrowed from aviation safety—is a situation in which an accident is narrowly avoided, such as when a driver brakes suddenly to avoid a crash (Arai et al., 2001). Telematics offers the possibility of recording this type of near-miss data in everyday transportation. Near-misses in the context of automobile insurance have recently been investigated by Guillen et al. (2020) in a sample of drivers with an OBD designed to collect information on driving indicators through continuous measurement. They found that risky driving behaviors such as driving fast or at night are associated with the occurrence of near-misses, but they did not apply this result to actual claims and they did not discuss insurance ratemaking. For their part, Gao et al. (2021) concluded that both classical actuarial risk factors and telematics car driving data are necessary to create the best predictive models for claims frequency.

In PHYD motor insurance policies also known as usage-based insurance (UBI), the premium is calculated on the basis of the customer's driving patterns, such as excess speed, hard acceleration, hard braking, or hard cornering—that is, potentially dangerous acts that we call near-miss events. Based on telematics information on drivers' performance, a risk score can be obtained which can be used to calculate the insurance premium. Arumugam and Bhargavi (2019) describe a PHYD payment approach based on risk scores at a theoretical level, but they do not discuss any real case study examples. Eling and Kraft (2020) analyze the use of telematics in insurance and conclude that the reduction of information asymmetries, as well as the possibility of carrying out more accurate risk pooling, improves the insurability of risks; however, they stress the need for further research on loss frequency and severity in the context of telematics data. Geyer et al. (2020) analyze a telematics data set with detailed information about driving behavior (speed, distance driven, and road type) in pay-as-you-drive (PAYD) contracts that are based on distance driven and road type. Weidner et al. (2016) discuss the problem of intractably large amounts of telematics information, also including driving context data, and the application of different methods for signal processing and driving pattern recognition. Although the authors are not able to assess the accident risk of any identified pattern, they conclude that given a comprehensive risk experience, these methods can be used for representing driving situations with the purpose of scoring driving behavior for pricing. Compared to them and to create a pricing structure, we compensate the lack of claims during the period when telematics information was collected with past claim history of insureds. We also provide details on the design of a pricing scheme based on telematics.

The literature in computer science and electrical engineering has also considered similar questions, specifically many contributions aim at identifying the context of driving events. In that sense, Qi et al. (2018) build a system called DrivAid that is able to detect three types of driving activities, namely, turns, hard brakes, and lane changes. They also extract context information from audio–visual signals in real time to develop a deep understanding of driving events. The advances in engineering are relevant to us because pricing based on near-miss information will only develop further if sensors are reliable and if stored telematics data are accurate.

Wüthrich (2017) was the first author to introduce pattern recognition and machine learning in actuarial modeling. All these studies agree that driving behavior data collected from telematics devices are helpful for auto insurance underwriting (see also Baecke & Bocca, 2017, Fan & Wang, 2017). Moreover, Ellison et al. (2015) conclude that personalized feedback alone is sufficient to induce significant changes in driving behavior, but that the largest reductions in risk are observed when drivers are also awarded a financial incentive to change. As part of a wide-ranging research project on behavior-based personalization in European insurance, Meyers and Van Hoyweghen (2020) report a case study carried out in 2016 aimed at tracking the driving behavior of 5000 participants through smartphone sensors in return for a 20% discount on their premium. However, they only recruited 243 participants in the study, and no clear evidence on the relation between driving style and loss ratios was found.

Many researchers whom we will mention below indicate that the occurrence of near-misses seems related to a higher risk of being involved in future accidents. However, near-misses have not been analyzed in the literature for the purposes of rating. In general, the existing contributions on telematics ratemaking refer to the measurement of general patterns related to driving style such as speed, acceleration, total distance, trip length, and so on, but not near-miss counts (see Lemaire et al., 2016 and the related contributions by Gao, Meng, et al., 2019; Gao & Wüthrich, 2019; Gao, Wüthrich, et al., 2019; Silva & Henriques, 2020). To the best of our knowledge, the combined dynamics of near-misses and claims frequency data have not been assessed with a view to designing an insurance ratemaking scheme. The paper by Denuit et al. (2019) proposes the use of multivariate mixed models to describe the joint dynamics of telematics data and claim frequencies. The approach proposed in that paper recognizes the *a posteriori* nature of telematics data and their variety among insured drivers. However, there are two major differences between Denuit et al. (2019) and what is presented here. First, they use *a posteriori* corrections that are based on both past claims and what they name "the signals," that is discretized telematics information. Their contribution is inspired in the bonus-malus, or experience rating principles in a multivariate setting, where counts coming from telematics information are integrated. Second, Denuit et al. (2019) do not look at the type of events that are considered in our current approach. They mainly work with claims, discretized mileage, and mileage driven under certain conditions, namely, night driving or urban driving. Denuit et al. (2019) implicitly look at yearly premium corrections because claims are usually compiled on a yearly basis and, although they admit that telematics are about to gain notoriety in the future, they do not discuss dynamic pricing mechanisms on a more frequent basis, that is weekly. However, their approach could be used with near-miss event counts.

The positive association of near-misses and accidents is well documented in the literature. Analyzing driving risk associated with near-crashes, Wang et al. (2015) found that the speed when braking and the potential crash type, among other factors, had the greatest influence on the driving-risk level of a near-crash. Ma et al. (2018) estimate a logistic regression model for the probability of accident and a Poisson regression for the number of accidents. Telematics and contextual driving information are used as explanatory variables. These authors find that vehicle mileage, hard brakes, hard starts, peak-time travel, and speeding are strongly correlated with higher accident rates; driving at a speed significantly different from that of traffic flow is also a relevant risk factor. Similarly to us, Ma et al. (2018) also use the past claim history of drivers (during the previous 3 years) and discuss premium implications, but they do not have information on smartphone usage. Stipancic et al. (2018) analyzed hard braking and accelerating events and compared them with historical crash data, and found that both maneuvers were positively correlated with crash frequency. Huang and Meng (2019) investigate the use of driving behavior variables for predicting the risk and frequency of claims in UBI products. They use logistic regression and four machine learning techniques for the risk probability models, and Poisson regression for the claim frequency model. They have five types of explanatory variables: traditional ratemaking factors, usage, travel habits, driving performance, and critical incidents. They carry out an empirical analysis with data from an insurance company in China. Compared to us, they have a sample of drivers larger than ours, but their drivers are observed during a shorter period of time. Their telematics information is quite similar to ours, but they do not include smartphone usage. The main difference is that they use driving performance variables and critical incidents as explanatory variables, while we mainly focus on the role of near misses and in our case, only the total distance traveled is included as a driving performance variable. Similarly to us, Huang and Meng (2019) conclude that driving behavior variables have a great potential in automobile insurance and ratemaking for UBI products. What makes the two contributions different is the objective. We focus on ratemaking and the role of near-misses, or incident, in dynamic pricing, whereas Huang and Meng (2019) confirm that the frequency of some types of near-misses is associated with the frequency of claims.

Focusing on three types of near-misses (accelerations, braking, and cornering), Guillen et al. (2020) found that age was significant for predicting near-miss events, but that gender did not have a significant effect. These authors also found that engine power was associated with a higher frequency of cornering and acceleration events; that nighttime driving presented a significant association with a lower risk of cornering events than daytime driving; that urban driving was associated with a higher frequency of braking events; and that excess speed increased the expected frequency of abnormal accelerations. Hynes and Dickey (2008) found that even the rate of change of acceleration had an influence on rear-impact vehicle collisions. We refer the reader to Eling and Kraft (2020; Table B1) where they can find an updated, exhaustive and systematic summary of existing model approaches in telematics insurance.

## 3 | METHODS: FREQUENCY MODELS AND BILLING SCHEMES

This section presents the methodology that has been used to design the billing scheme. First, we describe the classical frequency model and second we introduce near-miss information to obtain a pricing scheme. The fundamental aim of near-miss telematics motor insurance is to reward safe driving. We believe that this can be achieved by linking safe driving to lower insurance prices.[2]

Let $Y_i$, with $i = 1, ..., n$ denote the number of claims corresponding to observation $i$ during a specific period of time, $T_i$, where $n$ represents the total number of observations and here, $T_i$ is the duration of policy $i$. Let $x_i = (x_{i1}, ..., x_{ik})$ represent the vector of $k$ exogenous variables or risk factors that are assumed to have an effect on the expected number of claims and which are collected when the policy is underwritten, so they are static. Usually, the first component, $x_{i1}$, equals 1 to have an intercept in the model. Let us assume that given $x_i$, the dependent variable $Y_i$ follows a Poisson distribution with parameter $\lambda_i$, which is a function of the linear combination of parameters and regressors, $\beta_1 x_{\mathrm{i1}} + \cdots + \beta_{\mathrm{k}} x_{\mathrm{ik}}$, where $\beta = (\beta_1, ..., \beta_{\mathrm{k}})$ represents the vector of unknown parameters to be estimated in the Poisson model. The linear predictor is usually denoted $x_{\mathrm{i}}'\beta$.

An offset is included in the model when the exposure to risk is not the same for all drivers. $T_i$ is also called the *exposure to risk* for individual $i$. In that case, the model can incorporate the risk exposure factor (or the offset variable $\ln T_{\mathrm{i}}$) as follows:

$$E(Y_{\mathrm{i}} \mid x_{\mathrm{i}}) = \exp(\ln T_{\mathrm{i}} + \beta_1 x_{\mathrm{i1}} + \cdots + \beta_{\mathrm{k}} x_{\mathrm{ik}}) = T_{\mathrm{i}} \cdot \exp(\beta_1 \mathrm{x}_{\mathrm{i1}} + \cdots + \beta_{\mathrm{k}} x_{\mathrm{ik}}), \tag{1}$$

where $\exp(x_{\mathrm{i}}'\beta)$ provides a model for the number of claims per unit of exposure. If the duration of contracts is homogeneous then $\mathrm{T_i}$ is constant and it can be droped from Model (1).

In our real case analysis, each individual in the telematics data set corresponds to a vehicle-week observation. Therefore, each observation contains the telematics and event information collected for a specific vehicle during a specific week, and driver $i$ is observed over $W_{\mathrm{i}}$ weeks. However, when we consult past information on claims the duration of contracts is not usually homogeneous. The past claims history available for a contract in an insurance company generally extends over more than 1 year. Then, the offset variable has to be appropriately rescaled to allow comparisons between the past claims data set and the weekly summaries of the telematics datasets.

### 3.1 | The role of risk exposure

Risk exposure must be carefully handled in the two sources of information: first, the historical records, and second, the incoming telematics data. We suggest that a telematics summary be available each week; accordingly, we accommodate our timeframe to weekly data.

Past claims history reflects the duration of the insurance contract and the reported claims. Indeed, someone who is covered for 2 years by a motor insurance policy is exposed twice as long as someone who is covered only 1 year. We will not include distance driven in the coverage period because the historical records of traditional insurance companies do not generally include this information.[3] So, we consider only contract time duration, $T_{\mathrm{i}}$, in the

[2] In fact, this is not a new concept in insurance: Traditional bonus-malus systems reward good drivers with no claims with a *bonus* rebate in the following year's premium, and penalize drivers with claims with an extra charge, or *malus*, in the following year's price.

[3] Note that in the case study we also consider that the distance driven in the historical coverage period can be estimated by the average weekly distance driven observed in the telematics observation period. The results are presented in the Appendix.

exposure to risk component of Model (1) to cope with unequal contract lengths for the policyholders sampled. If $T_i$ is expressed in weeks, then a policy that covers 1 year, has $T_i$ approximately equal to 52.25 weeks.

On the other hand, we want to use telematics data to estimate the driver's behavior behind the wheel. We can imagine that this telematics analysis might be a sort of screening, in which drivers agree to be monitored by the insurer to assess their driving patterns and estimate the claims risk. So, an OBD telematics system would normally be installed in the vehicle and data would be recorded for a fixed and short period of time: for example, a few weeks. Then, we may take into account distance driven during the telematics observation period, so that the number of near-miss events that are observed can be related to the distance driven. In fact, each week, we may consider the absolute frequency observed of near-misses, that is, frequency counts, or alternatively we may take into account the distance driven that particular week, depending on whether or not we consider that the average weekly distance driven does change over time and, consequently counts per kilometer can be considered in our models.

### 3.1.1 | Near-misses per kilometer driven

Let us start with the traditional claims frequency model, like (1), $E(Y_i \mid x_i) = T_i \exp(x_i'\beta)$, where $T_i$ is the contract duration in weeks and $Y_i$ refers to the frequency of claims occurring during the contract coverage. We have one observation per driver. If we include near-miss information in the linear predictor, then one specification including the near-miss telematics data can be written using the relative frequencies of near-misses per kilometer, as follows $z_i'\theta = \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \alpha_1 E_{1i} + \dots + \alpha_s E_{si}$, where $\theta = (\beta, \alpha) = (\beta_1, \ldots, \beta_k, \alpha_1, \ldots, \alpha_s)$ is the vector of parameters to be estimated and $E_i = (E_{1i}, \ldots, E_{si})$ represents the near-miss average for individual $i$, and we denote the vector of all covariates as $z_i' = (x_i, E_i)$. Here, we record telematics information for each driver over a number of weeks $W_i$, sum the total number of near-miss events, and divide them by the total distance driven during $W_i$ weeks to obtain the near-miss average.

Our unit of analysis is the driver. On the one hand, we have information on the past claims history and general static characteristics of the policy and, on the other hand, we include average near-misses per distance driven, which we obtain from the telematics observational window. So, our approach is equivalent to extending the model with additional regressors. Model (1) becomes.

$$E(Y_i \mid x_i) = T_i \exp(x_i'\beta + E_i'\alpha) = T_i \exp(x_i'\beta)\exp(E_i'\alpha). \tag{2}$$

The claims frequency model has three components: $T_i$, which indicates weeks at exposure for the past-claims information, $\exp(x_i'\beta)$, which measures the influence of exogenous factors that are available in the insurance policy and $\exp(E_i'\alpha)$, which measures the influence of weekly near-miss information relative to the distance driven.

While Model (2) is useful for understanding the driver's behavior, its main limitation is that the incoming weekly telematics information is smoothed in the near-miss average for driver $i$ and we cannot directly see the impact of each particular individual near-miss event.

### 3.1.2 | Weekly near-miss counts

If we want to include near-miss events directly as counts, we ignore the distance driven and use observation time as risk exposure. For example, we may assume that drivers have repetitive behavior over time and tend to have constant exposure to risk. In this case, the rates would be proportional to

$$E(Y_i \mid x_i) = T_i \exp(x_i{}'\beta)\exp(E_{it}^{*}{}'\alpha), \tag{3}$$

where $T_i$ indicates the exposure to risk in weeks for the claims records for driver $i$, and $E_{it}^{*}{}'$ is the near-miss vector of absolute frequencies, that is, a count data vector for driver $i$ and Week $t$. The problem here is that the number of observations from the past claims is one per driver, while the number of telematics observations is $W_i$ per driver.

One way to overcome the problem of the unequal number of data points in the claims records and in the telematics data set is to replicate $W_i$ times the past claims information, so that $Y_i = Y_{it}$ for all $t$, and similarly for the static regressors $x_i{}'$. Then, we may analyze driver-week data. This means that if we have $n$ drivers each observed over $W_i$ weeks, then the final data set has $\sum_{i=1}^{n} W_i$ observations and we initially treat them as independent observations. Note that the positive correlation between observations of the same vehicle impacts the model inference causing an overestimation of standard errors, but our aim is primarily to produce predictive models. Alternatively, a Poisson panel regression can be specified, and these results are presented for our case study in the Appendix.

## 3.2 | Design of a PHYD motor insurance rating scheme with near-miss telematics data

Our pricing model stems from the calculation of the basic pure premium, where the pure premium price is calculated as the product of the expected cost of a claim, for which we take a general average, $\bar{C}$, times the expected number of claims.

With Model (3), a weekly premium ($T_i = 1$) is proportional to the product $\bar{C}\exp(x_i{}'\beta)\exp(E_{it}^{*}{}'\alpha)$. As we can see, when there are no near-miss events, the premium is $\bar{C}\exp(x_i{}'\beta)\exp(0) = \bar{C}\exp(x_i{}'\beta)$, a basic pure premium, which we call $P_{ibase}$, whereas each additional near-miss $j$ multiplies the premium by $\exp(\alpha_j)$, which is a relative increase in the premium (if $\alpha_j$ is positive) or a decrease (if $\alpha_j$ is negative). In this way, we can introduce a ratemaking structure in which the premium is dynamically updated when near-misses occur.

We expect near-misses to have positive coefficients so that the premium increases every time that the frequency of near-misses increases. Model (3) works on a scheme where the fitted insurance premium is recalculated whenever new information on counts reaches the insurance company through a telematics gate. If the insurance company prefers to establish a charge per kilometer, then distance-driven information could be included as an extra regressor of the telematics part or as an additional offset.[4] We do not include distance driven in (3) because we

[4]Following Boucher et al. (2013, 2017), we believe that premium rates should not be proportional to distance driven, because there is a natural learning curve. This means that intensive drivers who drive more kilometers are more experienced than novel or occasional drivers; therefore, each additional kilometer should cost slightly less for experienced drivers than for occasional drivers (see also, Guillen et al. 2019). Recently Boucher and Turcotte (2020) have investigated again the relationship between distance driven and claim frequency and they have shown that an approximately linear relationship can be derived once additional information is included. Specifically, the authors conclude that the learning curve that we seem to observe when looking at total distance driven over 1 year can actually be explained by other factors, such as more frequent use of the highway, higher proportion of driving outside rush hours, and so on. They conclude that, for each driver, independently of their driving risk per kilometer, the risk of an accident increases by approximately 1/15,000 for each additional kilometer driven.

think that it tends to be quite stable over time, it is correlated with the number of near-misses, and price tranches can be established depending on the average weekly distance driven by policyholders, in the same way, that portfolio segments or risk classes can be used combined with our pricing scheme.

The near-miss telematics ratemaking scheme that we propose is the product of a weekly premium $P_{\text{ibase}}$ that depends on static characteristics collected at policy issuance times a factor that takes into account the influence of each additional near-miss event, namely the factor contained in vector $E^*_{\text{it}}$. We can use the following approximation for the weekly premium to interpret the coefficients, $\alpha$, as percent increases per near-miss event of the base rate $P_{\text{ibase}}$, but we can also approximate a linear rate that would penalize each additional near-miss:

$$\bar{C}\exp(x_{\text{i}}{}'\beta)\exp(E^{*\prime}_{\text{it}}\alpha) = P_{\text{ibase}}\exp(E^{*\prime}_{\text{it}}\alpha) \cong P_{\text{ibase}}(1 + E^{*\prime}_{\text{it}}\alpha) \le P_{\text{ibase}} + E^{*\prime}_{\text{it}}\alpha_{\max}, \tag{4}$$

where $\alpha_{\max} = \max\limits_{1\le \text{i}\le \text{n}}(\alpha \text{P}_{\text{ibase}})$.

Note that $\alpha_{\max}$ depends on the maximum value of $P_{\text{ibase}}$. In practice, to determine $\alpha_{\max}$ a reasonable threshold for $P_{\text{ibase}}$ could be, for example, three times the average of $P_{\text{ibase}}$.

The availability of near-miss information gives insurers with telematics equipment the opportunity to promote the use of discounts and penalizations more often than the yearly adjustments that are produced by bonus-malus systems (also known as *experience rating* mechanisms, meaning that the price adjusts to the number of past claims). In a sense, we propose to include near-miss records as a way of reflecting the evidence of dangerous driving by means of the frequency of near-miss counts. This corresponds to the last term in (4).

## 4 | CASE STUDY

The database used in this article consists of a sample of 641 drivers with an insurance policy that includes a telematics recording system. Telematics information is available for these policyholders for the period between the 30th week of 2016 until the 30th week of 2019, but not all drivers were observed during the whole period. None of the drivers reported any claims during the period in which their corresponding telematics information was collected. In total, there are 7570 vehicle-week observations in the telematics data set.

### 4.1 | Data description

The list of variables and their descriptions are provided in Table 1. All drivers in our sample carried an OBD in their car, most of them for several months. Bolderdijk et al. (2011) reported that after 3 months drivers tend to forget that their driving behavior is being monitored, however, we did not analyze this change of behavior here.

Given the small observation period per driver, it is unlikely to observe a claim in any given driver. A typical annual frequency for "at fault" claims is below 5%, in most developed countries, which means that, since we have observed 7570 vehicle week observations, we should be expecting to observe about seven claims in the telematics observation period. Note that drivers were aware that they were monitored and it is well known that this fact may induce prudence on the wheel. Therefore, the information on claims by drivers in our telematics data set is taken from the past claims history made available by the insurance company. The claims

**TABLE 1** List of variables in the telematics/claims data set

| Variables | Description |
|---|---|
| Near-miss event information | |
| EAclr1, EAclr2, EAclr3 | Acceleration event (three intensities: 1, 2, and 3) |
| EBrak1, EBrak2, EBrak3 | Braking event (three intensities: 1, 2, and 3) |
| EPhone | Smartphone usage event (in seconds) |
| Other telematics information | |
| DistThous | Total distance traveled during the week (in 000 s Km) |
| Claim information | |
| ExpoT | Exposure time (in weeks) |
| NumT | Total number of at-fault claims during the exposure period. |
| Vehicle information | |
| EngineCapacity | Engine capacity of the vehicle (in 000 s cc). |

file includes the total number of claims at fault (NumT), and the exposure period measured in weeks (ExpoT). Additionally, we also know the level of severity of these claims, classified in four categories: low, medium, high, and extreme.

The telematics information was collected at weekly intervals and included for each vehicle and week total distance traveled in 000 s (DistThous). Preprocessing and curation of data leading to weekly summaries were performed by the telematics manufacturer and no major issue was reported. Additionally, for each vehicle and observed week, the number of near-miss events of various kinds was computed, including acceleration (EAclr) and braking (EBrak). For each event type, we know the intensity level (1 = moderate, 2 = medium, and 3 = dangerous). Note that the data were provided to us by a telematics manufacturer with limited information on these three classes of severity which was considered proprietary, and this is a limitation. The identification of each near-miss is based on a severity score that lies between [0,10] (following the same definitions as in Guillen et al., 2020). For example, in the case of acceleration events, the intensity score level takes into consideration the difference between the maximum acceleration reading and the acceleration detected in the first reading above the acceleration event detection threshold, set at 6 $m/s^2$ in accordance with the previous literature: see, for example, Hynes and Dickey (2008), where 5.7 $m/s^2$ was considered as the threshold for a low peak acceleration event during rear-end impacts. We calculate the ratio between this difference and the corresponding timestamps of the latter readings. A transformation of this ratio multiplied by 10 yields the final severity score, which lies between [0,10]. Acceleration is also used to determine the severity of braking events, given that negative acceleration can essentially be considered as deceleration. Smartphone usage (EPhone) while driving is also recorded. Regarding the variables that are traditionally used in motor insurance pricing, only engine capacity and engine horsepower were available to us. In fact, they are redundant because they both measure the vehicle's engine power. Engine capacity was included in the final model, while engine horsepower was finally excluded.

Table 2 shows some descriptive statistics and also includes the sum of events by type, such as EBrak (total number of brakes, i.e., sum of EBrak1, EBrak2, and EBrak3), EAclr (total number of acceleration events, NumT [total number of at-fault claims]) and total exposure in weeks. Additionally, Figures 2–4 show the histogram of EBrak, EAclr, and EPhone. Due to the

**TABLE 2** Descriptive statistics in the telematics and claims data sets

| Variable | Mean | *SD* | Minimum | Maximum |
|---|---|---|---|---|
| EBrak1 | 1.8629 | 6.3567 | 0 | 93 |
| EBrak2 | 0.6764 | 2.7104 | 0 | 33 |
| EBrak3 | 0.1703 | 1.1368 | 0 | 29 |
| EBrak | 2.7095 | 8.6934 | 0 | 119 |
| EAclr1 | 1.3931 | 7.1128 | 0 | 202 |
| EAclr2 | 0.1180 | 0.7488 | 0 | 20 |
| EAclr3 | 0.1655 | 1.1916 | 0 | 30 |
| EAclr | 1.6766 | 7.9614 | 0 | 219 |
| EPhone | 16.0008 | 77.7907 | 0 | 4150 |
| DistThous | 0.1523 | 0.1865 | 0.0010 | 2.7230 |
| EngineCapacity | 1.8383 | 0.7328 | 0.4250 | 6.2550 |
| NumT | 0.0764 | 0.4352 | 0 | 6 |
| ExpoT | 287.3532 | 53.5283 | 52.2857 | 545.8571 |

large frequency of zeroes we represent only the positive observations. The data present a long right tail, so we also decided to limit the representation up to a maximum value, specifically 50 for EAclr and EBrak, and 300 for EPhone. Note that EAclr has 83.66% of zeroes, and 0.62% are equal or greater than 50. EBrak has 80.91% of zeroes, and 0.82% are equal or greater than 50, and finally, EPhone has 79.78% of zeroes and 0.65% are higher than 300.

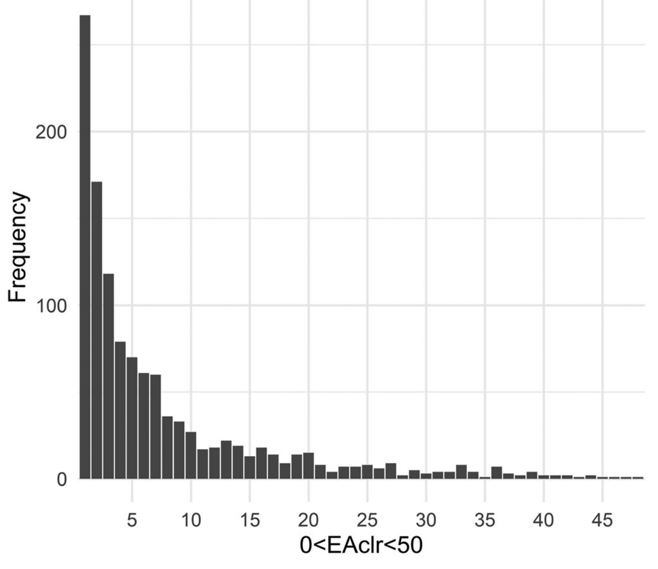


**FIGURE 2** Histogram of EBrak (the number of braking events)

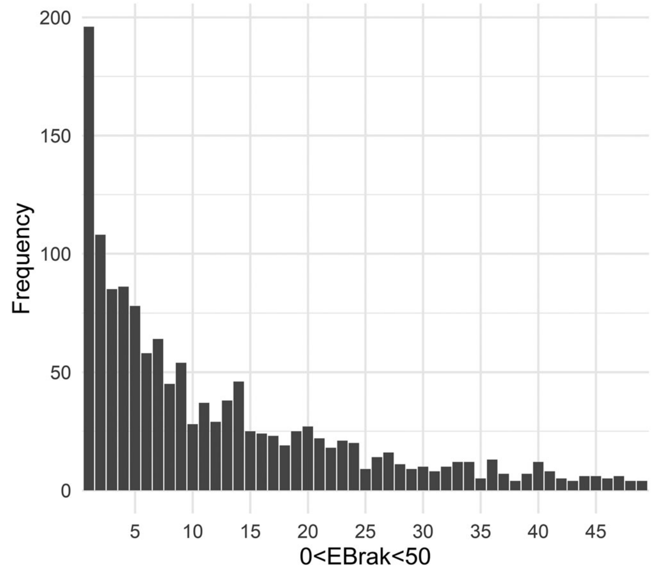


**FIGURE 3** Histogram of EAclr (the number of acceleration events)

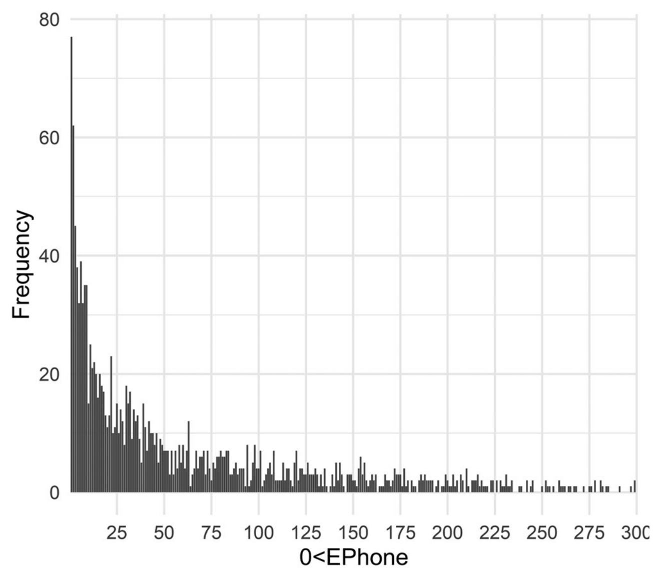


**FIGURE 4** Histogram of EPhone (seconds of smartphone usage)

The most frequent event was smartphone usage; drivers recorded an average of 16.00 s of smartphone usage while driving per week. The second most frequent event was braking, with a weekly average of 2.71. Finally, acceleration events were the least frequent events, with an average of 1.68. For each event type, events of intensity Level 1 were the most frequent. Regarding telematics variables, the average distance traveled per week was 152.3 km. The average engine capacity was 1838 cc. The claims history of the drivers in the telematics sample reflected an average weekly number of claims equal to 0.0004, which corresponds to a yearly claim frequency of 0.0316. This is a rather low frequency because only claims at fault were considered in the design of our ratemaking structure.

## 4.2 | Results

We estimated a Poisson regression model like (3) where the number of claims is the dependent variable. The exogenous variables were: braking events, acceleration events, smartphone use, and engine capacity. Table 3 shows the results of the Poisson model.

In Table 3, we observe that acceleration events of Intensity 1 (EAclr1) have a significant effect ($p = .0019$) and are associated with a lower claim frequency because the coefficient is negative (−0.0825). On the other hand, acceleration events of Intensity 2 (EAclr2) are significantly associated with a higher claim frequency with a positive coefficient (0.3069; $p = .0162$). Acceleration events of Intensity 3 (EAclr3) do not have a significant effect in the model ($p = .8072$). We interpret that a moderate level (Intensity 1) of acceleration avoids accidents, while Intensity 2 is associated with aggressive driving and thus, with more claims. The conclusion regarding braking events is similar: braking events of Intensity 1 and Intensity 3 are associated with a higher claim frequency with coefficients equal to 0.0268 and 0.0984, respectively and $p$-values lower than 1%, while braking events of Intensity 2 are associated with a lower claim frequency. This means that braking events of Intensity 2 seem to be associated with avoiding at-fault accidents. Smartphone usage has a significant effect at the 10% significance level ($p = .0776$). Finally, EngineCapacity has a positive and significant effect on the claim frequency.

Given that some near-miss frequency variables have a negative coefficient, we conclude that some braking or accelerating events are not necessarily bad. A driver who brakes whenever necessary is probably a good driver who is able to react swiftly in situations of danger. A similar conclusion can be reached for moderate acceleration events. Regarding the existence of three severity levels of near misses (rather than collapsing them in one

**TABLE 3** Parameter estimates of the Poisson model of the weekly rate of at-fault claims for the telematics and claims data set

| Parameter | Estimate | Standard error | *p*-Value |
|---|---|---|---|
| Intercept | −8.0637 | 0.0673 | <.0001 |
| EAclr1 | −0.0825 | 0.0265 | .0019 |
| EAclr2 | 0.3069 | 0.1277 | .0162 |
| EAclr3 | 0.0095 | 0.0390 | .8072 |
| EBrak1 | 0.0268 | 0.0086 | .0018 |
| EBrak2 | −0.4966 | 0.0770 | <.0001 |
| EBrak3 | 0.0984 | 0.0336 | .0034 |
| EPhon | 0.0004 | 0.0002 | .0776 |
| EngineCapacity | 0.3644 | 0.0287 | <.0001 |

*Note*: The Akaike information criterion equals 7345.00 and the Bayes Information criterion equals 7407.39. The pseudo-$R^2$ equals 21.83%.

single category), we would like to remark that it enables to address pricing by considering the fact that a higher frequency of near-misses does not necessarily correlate with a higher frequency of claims. We have seen that a moderate amount of near-misses may indicate that the driver has good reflexes to actually avoid accidents. This is why we do believe that a certain level of near-miss counts could perhaps not be penalized, but more data are more detailed on the definition of class severity of near-miss events is needed to analyze this question.

In addition, there is a strong positive correlation between the number of pricing events and the number of kilometers driven. Therefore, if the number of kilometers was to be considered in the pricing model, the marginal impact of pricing events (brakes, accelerations, etc.) would be low, as it would already be captured by the total number of kilometers. Indeed the number of kilometers driven and near-misses are very much correlated, though not homogeneously for everyone. This is the main reason why insurers tend to prefer UBI, based on a comprehensive view of driving patterns, rather than PAYD which only pays attention to distance driven. We believe that our approach encompasses the two worlds. In fact, the price paid depends both on the distance and on the near-misses. There are natural and straightforward ways to weight the importance of these two components and that would be a mission for insurers to decide what combination of a pricing scheme they would prefer to offer. For example, an insurer could decide to compose a final UBI pricing scheme where only 10% of the price depends on distance, while 90% depends on near-misses. Another insurer could set a price mechanism where 90% of the price would depend on distance, while only 10% would depend on near-misses. In general, one could fix a linear combination of both components, where a pure PAYD scheme would arise naturally as a particular case when distance takes 100% of the pricing weight and near-misses have a 0% weight.

A limitation of our model is that we replicate the past claim information across person-specific observations and treat them as independent observations. Even if we focus on prediction, we need to have valid information on the statistical significance of our estimates. Therefore, we have introduced a correction for the standard errors to cope with the panel effect. Specifically, we have introduced fixed effects and we have also tried a week trend. The results are presented in Table A1. Our conclusions do not change. The model presented in Section 3.1.1 aggregates all telematics information into one single row per driver and then combines the data to produce a model for the number of claims. The results again produce similar conclusions and are available from the authors.

Finally, we would like to remark another limitation of our analysis. We do not have data on the distance driven during the period in which claims were made. We could make the assumption that an individual drove as much back then as he/she drives during the period in which we collect telematic information. In that case, we could estimate a model that includes the logarithm of the distance traveled during the coverage period. The results of this alternative model are presented in Table A2. The logarithm of the distance, as expected, has a positive significant effect, but the rest of the variables in the model do not change after introducing this new variable. An alternative approach to our model would be to re-estimate the model taking into account only events with intensity higher than 2 (which for both braking and acceleration have a positive correlation to claims). In that case, only EBrak has a significant (positive) effect. The results are shown in Table A3. Given that we did not observe distance driven in the claims period, we have only provided the basic model results including distance.

**TABLE 4** Weekly breakdown of a total bill per week

| Week | Distance driven (km) | $a$ = Number of near-miss brakes | $b$ = Number of near-miss accelerations | $c$ = Minutes of smartphone use | $d$ = Cost of near-misses (Eur) | $e$ = Bill per week (Eur) |
|---|---|---|---|---|---|---|
| 1 | 30 | 0 | 0 | 0 | 0.00 | 1.95 |
| 2 | 73 | 0 | 0 | 2 | 0.37 | 2.32 |
| 3 | 104 | 2 | 2 | 2 | 6.59 | 8.54 |
| 4 | 260 | 6 | 2 | 1 | 9.40 | 11.35 |
| 5 | 705 | 19 | 4 | 21 | 27.51 | 29.46 |

*Note*: Pure premium in motor insurance as a function of near-miss events for a driver of a car with engine capacity 1769 cc. Basic weekly rate 1.95 Eur.

Total bill for 5 weeks: 53.61 Eur.

$e = 1.95 + d$.

$d = 0.75 \times a + 2.36 \times b + 0.18 \times c$.

## 4.3 | Examples of a billing process for near-miss telematics motor insurance

A step-by-step elaboration of the billing mechanism follows easily from implementing the model results in (4) and additional details are provided in Appendix B. To show how *near-miss billing* can be performed, let us consider an example driver in our real case study data set who had no claims in the past 5 years and who was observed from Week 20 in 2018 to Week 11 in 2019. During these 44 weeks this driver, whose car had an engine capacity of 1769 cc,[5] did not report any claims. Our billing scheme would start by charging this driver a flat fixed rate which could be based solely on the type of car, that is, on its engine capacity, and possibly other standard characteristics such as driving zone, age of the car, and so on. Then, the insurer's near-miss telematics rate would proceed by adding extra costs for near misses or phone use. We will also explore a second billing option: A slightly higher fixed basic rate than in our initial proposal, with the addition of extra charge for each type of near-miss event only if the near-misses exceed a certain level (e.g., two per week). We will explore the ratemaking examples in the following two subsections.

### 4.3.1 | Basic rate plus additional cost of near misses

We illustrate our approach via one concrete insured from our portfolio. In this section, we want to illustrate the practical outlook of our new approach. The insurer would bill a flat pure rate of 1.95 Eur per week, which is equivalent to a yearly rate of 101.84 Eur. Each near-miss brake would cost 0.75 Eur, each near-miss acceleration would cost 2.36 Eur, and each active minute of mobile phone while driving would cost 0.185 Eur. Additional insurance costs should be added to account for solvency, marketing and administration, among others.

Table 4 displays an illustration of the price paid in 5 consecutive weeks, according to the driver's experience. The first week the example driver did not have any near-misses and so he

[5]This driver had an average number of near-miss events per week equal to 2.7 hard brakes and, 1.1 hard accelerations, and his weekly average phone usage during driving was only 2.7 s. His total distance traveled during the observation window was 13,400 km, which is an average of about 300 km/week.

paid the basic rate equal to 1.95 Eur. In Week 2, he used the smartphone for 2 min, but the use was minimal and was barely penalized. In Week 5, the number of brakes and accelerations increase, as did phone use, adding 27.51 Eur to the basic rate and thus giving a total bill of 29.46 Eur for that week. Over this 5-week period, the policy-holder would pay a total sum of 53.61 Eur. This is of course a pure premium based on near-miss information, and so expenses and insurer-added loadings should be added on top. In addition to the near-miss scheme, a *per kilometer* rate could also be considered here as in a UBI scheme, but we do not include this here and so the distance driven column had no impact on the bill for each week.

The practical implementation illustrated in this subsection of our new near-miss approach to telematics is of course just one out of many possibilities. In the next subsection, we illustrate our approach via a slightly difference market concept.

### 4.3.2 | Basic rate with a reward for safe driving and the additional charge for near misses

In the previous subsection, near-miss events were charged at some given cost on top of some low basic fee. In this subsection, we reverse the order of payment and start with a high basic fee with a bonus or discount when the absence of near-miss events is recorded.

In the scenario where insurers explicitly want to reward safe driving, they would have to bill a basic rate higher than the previous example: For example, 6.66 Eur per week, which is the price for an average driver. They then apply a discount (−5.65 Eur) if no near-miss records are observed each week. Intermediate discounts may exist depending on the type of near-miss until their frequency exceeds a threshold (up to 2 for accelerations and up to 1 for braking and smartphone use). If the counts are larger than 2, then the penalizations are twice those in the cost scheme for the example shown in Table 4 in the previous subsection.

Table 5 displays the price that the driver in the above section would pay over the 5-week period. In the first couple of weeks, since the counts of near-misses are below the threshold, the insured policyholder would obtain a full discount of 5.65 Eur in the first week, so his weekly bill equals 6.66 – 5.65 = 1.01 Eur. In the second week, the discount is 5.29 Eur because

**TABLE 5** Weekly bill of pure premium in motor insurance as a function of near-miss events for a driver of a car with engine capacity 1769 cc)

| Week | Distance driven (km) | $a$ = Number of near-miss brakes | $b$ = Number of near-miss acceleration | $c$ = Minutes of smartphone use | $d$ = Cost of near-misses (Eur) | $e$ = Total weekly bill (Eur) |
|---|---|---|---|---|---|---|
| 1 | 30 | 0 | 0 | 0 | −5.65 | 1.01 |
| 2 | 73 | 0 | 0 | 2 | −5.29 | 1.37 |
| 3 | 104 | 2 | 2 | 2 | 0.93 | 7.59 |
| 4 | 260 | 6 | 2 | 1 | 9.00 | 15.66 |
| 5 | 705 | 19 | 4 | 21 | 54.94 | 61.60 |

*Note*: Basic weekly rate (6.66 Eur) minus discounts for safe driving, or plus penalizations for near misses.

Total bill for 5 weeks: 87.23 Eur.

$e = 6.66 + d$.

$d$ = if $a > 2$, $1.5(a)$, $-0.75(1-(a))$; if $b > 2$, $4.71(b)$, $-2.36(2-(b))$; if $c > 2$, $0.36(c)$, $-0.18(1-(c))$.

of the minimal smartphone use and the bill for that week would be 6.66 – 5.29 = 1.37. In the third week, the near-misses would not exceed the threshold count for braking and acceleration, but overall the driver receives a lower discount than before and he would be penalized 0.93 Eur. So, his total bill for the third week would be 6.66 + 0.93 = 7.59. Overall, these 3 weeks turn out better under this billing scheme than in the example shown in Table 5. From that week onwards all near-misses incur a penalty twice that in the example in Table 5, so the extra cost added to the flat rate is 9.00 Eur for Week 4, and the bill would come to 6.66 + 9.00 = 15.66 Eur. For Week 5 the extra cost would be greater, 54.94 Eur, and the bill would come to 61.60 Eur.

There is a substantial difference between the rating schemes displayed in Tables 4 and 5. Overall the second scheme rewards safe driving more than the first, while dangerous driving increases the sum of the weekly bills considerably. For this particular example in Table 5, the total sum of the pure premium paid after 5 weeks is 87.23 Eur.

## 5 | CONCLUSIONS

We have introduced near-miss telematics to insurance ratemaking via a recently compiled data set with high-quality telematics information but with only historical claims data. This is both a strength and a weakness of our study. The weakness is that the prediction would be more precise if claims were recorded using only the high-quality telematics data. However, this is also its strength, because our approach using historical claims information from customers is extremely useful for newcomers to the telematics business; one can start with dynamic near-miss telematics pricing after a few weeks of data collection when a portfolio of customers with previous claims records in the company is available. Our current study, therefore, introduces a new concept of near-miss telematics and at the same time introduces a way to address the challenging period of transition from classical insurance rating to near-miss telematics ratemaking. In that sense, there are many ways to bill near-misses and we selected two of the main strategies after interviewing several stakeholders. That is, we have concentrated on near-misses implying charges (driver pays per near-miss) and we have also contemplated that the absence of near-misses induces discounts, which seems to be the preferred way to go. Unfortunately, we cannot report how consumers will react to near-miss ratemaking as this has not been implemented in practice. We do not have references to compare to other approaches. For PAYD schemes we have not found evidence of drivers changing their behavior, that is driving less, due to the adherence to a PAYD contract. We are not able to discern whether near-miss telematics has any additional benefit over a distance-based pricing scheme, and we are unable to say from our empirical analysis whether drivers adopting telematics schemes will in general change their behavior in the long term as a consequence of the impact on the price of their usage-based insurance ratemaking. (see, Meyers & Van Hoyweghen, 2020, for an extensive economic experiment discussion).

Regarding claim severities, our approach has some limitations. We do have information on the severity of past claims, but we do not introduce a model for severities. We follow the classical ratemaking principle based on a frequency model times average cost. We wanted to keep the approach as simple as possible and we wanted to avoid claims cost data which is generally known to be quite noisy. Another limitation of our study is that we do not have data on the distance driven during the period in which claims were made. Nevertheless, we saw that inferring this distance from the available telematics information does not change much the conclusions of our analysis. An additional limitation comes from the fact that some levels of intensity of the risk behavior are actually associated with significantly lower numbers of claims. The interpretation could be that some near-misses are good, meaning that the driver is able to react promptly to avoid crashes.

Of course, it would be extremely interesting to provide analytical results for an extended data set based on years of near-miss telematics data. No such data set exists at this moment in time. The near-miss telematics approach needs to be implemented first using a transition data set from classical insurance to near-miss telematics, as illustrated in this paper. Therefore, while the approach of this paper seems most suitable at a time when no full near-miss telematics data exist, it would be of great interest to analyze a mature near-miss data set considered over several years with a significant number of customers. However, even with the simpler transition data set considered in this paper, we obtain very promising results, and we consider this paper to be a proof of concept that our near-miss telematics pricing could indeed be introduced at this point in time. Choosing the right structure of penalization of near-misses is a question that will remain in the area of each insurer's ratemaking strategy. Maybe some would decide to penalize only severe near-misses, whereas others may have much more refined systems, even nonlinear near-miss penalization which would match what we have found in the model results. However, the impact of near-miss telematics ratemaking on driving behavior and, ultimately, on claims is now an open question and we believe that further theoretical and empirical research will need to be carried out in the future and specially if near-miss ratemaking schemes are implemented in practice.

## ACKNOWLEDGMENTS

Montserrat Guillen and Ana M. Pérez-Marín thank Fundación BBVA Big Data grant and Spanish Ministry of Science and Innovation grant PID2019-105986GB-C21, Montserrat Guillen thanks ICREA Academia. The authors also thank the editors and two anonymous reviewers for providing feedback and helping us.

## ORCID

*Montserrat Guillen* http://orcid.org/0000-0002-2644-6268
*Jens Perch Nielsen* https://orcid.org/0000-0002-2798-0817
*Ana M. Pérez-Marín* https://orcid.org/0000-0002-8318-0603

## REFERENCES

Arai, T., Nishimoto, T., Ezaka, Y., & Yoshimoto, K. (2001). Accidents and near-misses analysis by using video drive-recorders in a fleet test. *Proceedings of the 17th International Technical Conference on the Enhanced Safety Vehicles*, Amsterdam. https://www.sae.org/publications/technical-papers/content/2001-06-0001/

Arumugam, S., & Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. *Journal of Big Data*, *6*(86), 1–21.

Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, *98*, 69–79.

Barry, L., & Charpentier, A. (2020). Personalization as a promise: Can Big Data change the practice of insurance? *Big Data & Society*, *7*(1), 2053951720935143.

Bolderdijk, J. W., Knockaert, J., Steg, E. M., & Verhoef, E. T. (2011). Effects of pay-as-you-drive vehicle insurance on young drivers' speed choice: Results of a Dutch field experiment. *Accident Analysis & Prevention*, *43*(3), 1181–1186.

Boucher, J.-P., Côté, S., & Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, *5*(4), 54.

Boucher, J. P., Pérez-Marín, A. M., & Santolino, M. (2013). Pay-as-you-drive insurance: the effect of the kilometers on the risk of accident. *Anales del Instituto de Actuarios Españoles 3ª Época*, *19*, 135–154.

Boucher, J. P., & Turcotte, R. (2020). A longitudinal analysis of the impact of distance driven on the probability of car accidents. *Risks*, *8*(3), 91.

Denuit, M., Guillen, M., & Trufin, J. (2019). Multivariate credibility modeling for usage-based motor insurance pricing with behavioural data. *Annals of Actuarial Science*, *13*(2), 378–399.

Eling, M., & Kraft, M. (2020). The impact of telematics on the insurability of risks. *Journal of Risk Finance*, *21*(2), 77–109.

Ellison, A. B., Bliemer, M. C. J., & Greaves, S. P. (2015). Evaluating changes in driver behaviour: A risk profiling approach. *Accident Analysis and Prevention*, *75*, 298–309.

Fan, C. K., & Wang, W. Y. (2017). A comparison of underwriting decision making between telematics-enabled UBI and traditional auto insurance. *Advances in Management and Applied Economics*, *7*(1), 1–2.

Gao, G., Meng, S., & Wüthrich, M. V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, *2019*(2), 143–162.

Gao, G., Wang, H., & Wuthrich, M. V. (2021). Boosting Poisson regression models with telematics car driving data. *Mach Learn*. https://link.springer.com/content/pdf/10.1007/s10994-021-05957-0.pdf

Gao, G., & Wüthrich, M. V. (2019). Convolutional neural network classification of telematics car driving data. *Risks*, *7*(1), 6.

Gao, G., Wüthrich, M. V., & Yang, H. (2019). Evaluation of driving risk at different speeds. *Insurance: Mathematics and Economics*, *88*, 108–119.

Geyer, A., Kremslehner, D., & Mürmann, A. (2020). Asymmetric information in automobile insurance: evidence from driving behavior. *Journal of Risk and Insurance*, *87*(4), 969–995.

Guillen, M., Nielsen, J. P., Ayuso, M., & Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis*, *39*(3), 662–672.

Guillen, M., Nielsen, J. P., Pérez-Marín, A. M., & Elpidorou, V. (2020). Can automobile insurance telematics predict the risk of near-miss events? *North American Actuarial Journal*, *24*(1), 141–152.

Huang, Y., & Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems*, *127*, 113156.

Hynes, L. M., & Dickey, J. P. (2008). The rate of change of acceleration: implications to head kinematics during rear-end impacts. *Accident Analysis and Prevention*, *40*(3), 1063–1068.

Lemaire, J., Park, S. C., & Wang, K. C. (2016). The use of annual mileage as a rating variable. *ASTIN Bulletin*, *46*(1), 39–69.

Ma, Y. L., Zhu, X., Hu, X., & Chiu, Y. C. (2018). The use of context-sensitive insurance telematics data in auto insurance ratemaking. *Transportation Research Part A*, *113*, 243–258.

Meyers, G., & Van Hoyweghen, I. (2020). 'Happy failures': Experimentation with behaviour-based personalisation in car insurance. *Big Data & Society*, *7*(1), 2053951720914650.

Qi, B., Liu, P., Ji, T., Zhao, W., & Banerjee, S. (2018). *DrivAid: Augmenting Driving Analytics with Multi-Modal Information*, 2018 IEEE Vehicular Networking Conference (VNC), Taipei, Taiwan. https://ieeexplore.ieee.org/document/8628415

Silva, M. I., & Henriques, R. (2020). Finding manoeuvre motifs in vehicle telematics. *Accident Analysis and Prevention*, *138*, 105467.

Stipancic, J., Miranda-Moreno, L., & Saunier, N. (2018). Vehicle manoeuvers as surrogate safety measures: Extracting data from the GPS-enabled smartphones of regular drivers. *Accident Analysis and Prevention*, *115*, 160–169.

Wang, J., Zheng, Y., Li, X., Yu, C., Kodaka, K., & Li, K. (2015). Driving risk assessment using near-crash database through data mining of tree-based model. *Accident Analysis and Prevention*, *84*, 54–64.

Weidner, W., Transchel, F. W. G., & Weidner, R. (2016). Classification of scale-sensitive telematic observables for riskindividual pricing. *European Actuarial Journal*, *6*(1), 3–24.

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal*, *7*(1), 89–108.

**How to cite this article:** Guillen, M., Nielsen, J. P., & Pérez-Marín, A. M. (2021). Near-miss telematics in motor insurance. *J Risk Insur*. 88, 569–589. https://doi.org/10.1111/jori.12340

## APPENDIX A: ADDITIONAL MODEL RESULTS

**TABLE A1** Parameter estimates of the Poisson model of the weekly rate of at fault claims for the telematics and claims data set with fixed effects

| Parameter | Estimate | Standard error | *p*-Value |
|---|---|---|---|
| Intercept | −27.1159 | 10538.89 | .9979 |
| EAclr1 | −0.0844 | 0.0286 | .0032 |
| EAclr2 | 0.2681 | 0.1349 | .0468 |
| EAclr3 | −0.0395 | 0.0407 | .3317 |
| EBrak1 | 0.0185 | 0.0121 | .1275 |
| EBrak2 | −0.6304 | 0.0833 | <.0001 |
| EBrak3 | 0.1071 | 0.0350 | .0022 |
| EPhon | 0.0003 | 0.0002 | .1597 |
| EngineCapacity | 0.3081 | 0.0303 | <.0001 |

*Note*: The Akaike's Information Criteria equals 7340.57 and the Bayesian Information Criteria equals 8297.17. The pseudo-$R^2$ equals 22.64%.

**TABLE A2** Parameter estimates of the Poisson model of the weekly rate of at fault claims for the telematics and claims data set (also including the logarithm of the distance travelled as explanatory variable)

| Parameter | Estimate | Standard error | *p*-Value |
|---|---|---|---|
| Intercept | −8.6026 | 0.1325 | <.0001 |
| EAclr1 | −0.0811 | 0.0266 | .0023 |
| EAclr2 | 0.2896 | 0.1281 | .0238 |
| EAclr3 | 0.0090 | 0.0386 | .8155 |
| EBrak1 | 0.0245 | 0.0087 | .0047 |
| EBrak2 | −0.5011 | 0.0769 | <.0001 |
| EBrak3 | 0.0822 | 0.0334 | .0137 |
| EPhon1 | 0.0003 | 0.0002 | .2379 |
| EngineCapacity | 0.3331 | 0.0295 | <.0001 |
| Log(Distance) | 0.1726 | 0.0357 | <.0001 |

*Note*: The Akaike's Information Criteria equals 7322.37 and the Bayesian Information Criteria equals 7391.69. The pseudo-$R^2$ equals 21.91%.

**TABLE A3** Parameter estimates of the Poisson model of the weekly rate of at fault claims for the telematics and claims data set (only near misses of intensity higher than 2)

| Parameter | Estimate | Standard error | *p*-Value |
|---|---|---|---|
| Intercept | −8.1588 | 0.0675 | <.0001 |
| EBrak3 | 0.0690 | 0.0166 | <.0001 |
| EngineCapacity | 0.3723 | 0.0288 | <.0001 |

*Note*: The Akaike's Information Criteria equals 7524.17 and the Bayesian Information Criteria equals 7544.96. The pseudo-$R^2$ equals 21.28%.

## APPENDIX B: PRICING WITH TELEMATICS NEAR-MISS COUNT DATA

Our proposed pricing model takes into account the number of brakes, the number of accelerations each week and smartphone use while driving. To simplify its practical implementation by insurance companies and by customers, the intensity of near-miss events may be disregarded at this stage. In addition, the premium can either be accommodated to the type of car (as a function of engine capacity), or it can be expressed as a uniform flat rate for all types of vehicles.

To calculate an example premium, we averaged the information on claim cost. In our data set, a low-intensity claim costs 500 Eur, a medium-intensity claim costs 5500 Eur, a high-intensity claim costs 30,000 Eur and an extreme-intensity claim costs 100,000 Eur. With these assumptions, the average at-fault claim cost in the claims record information in the sample was 3250 Eur. Note that we calculate pure premiums, and additional costs should be added by the insurer. The loadings to be added to the pure premium are general expenses, marketing, administration, and solvency.

Using the results of the model presented in Table 3, for an EngineCapacity equal to 1769 cc (see example in Section 4), and an average claim cost of 3250 Eur, we have that the weekly flat rate is approximately 1.95 and the additional cost of each braking event, where the additional cost of 0.75 Eur per each braking event is obtained by calculating $\alpha_{\text{maxbraking}}$ and three times the average EngineCapacity has been used. Proceeding in the same way, we can easily calculate the additional cost per each acceleration event, and the result equals 2.36. Similarly, the additional cost per each minute of smartphone usage equals 0.18.