The Geneva Risk and Insurance Review (2024) 49:75–126
https://doi.org/10.1057/s10713-024-00098-5

ORIGINAL ARTICLE

Big data, risk classification, and privacy in insurance markets

Martin Eling[1] · **Irina Gemmo**[2] · **Danjela Guxha**[1] · **Hato Schmeiser**[1]

Published online: 19 June 2024

Abstract
The development of new technologies and big data analytics tools has had a profound impact on the insurance industry. A new wave of insurance economics research has emerged to study the changes and challenges those big data analytics developments engendered on the insurance industry. We provide a comprehensive literature review on big data, risk classification, and privacy in insurance markets, and discuss avenues for future research. Our study is complemented by an application of the use of big data in risk classification, considering individuals' privacy preferences. We propose a framework for analyzing the trade-off between the accuracy of risk classification and the discount offered to policyholders as an incentive to share private data. Furthermore, we discuss the conditions under which using policyholders' private data to classify risks more accurately is profitable for an insurer. In particular, we find that improving the accuracy of risk classification, if achieved by requiring the use of private data, does not necessarily provide an incentive for insurers to create more granular risk classes.

Keywords Big data · Digitalization · Privacy costs · Risk classification

JEL Classification D43 · D81 · D82 · G22 · I13 · O31

1 Introduction

In recent years, advancements in big data, machine learning, and artificial intelligence (AI) have profoundly reshaped the insurance industry, ushering in a new era for insurance economics. Technological advances transform various aspects of insurance, from risk assessment to customer service. For instance, the

✉ Martin Eling
martin.eling@unisg.ch

[1] University of St. Gallen, St. Gallen, Switzerland

[2] HEC Montréal, Montreal, Canada

increased availability of detailed data, coupled with efficient data collection and analysis tools, enhances risk evaluation and cost estimation. This, in turn, benefits policyholders by mitigating issues like adverse selection and moral hazard. Additionally, AI contributes to improved service quality by providing better insurance services and streamlining claims management.

This paper contributes to the existing literature on big data, risk classification, and privacy considerations in insurance markets by providing a comprehensive review of the relevant research and presenting an application with respect to risk classification accounting for privacy costs. We discuss the impact of big data, machine learning, and artificial intelligence on risk classification in insurance by providing an overview of the literature on changes in the risk landscape of insurers and the implications for insurance market dynamics. Starting with seminal contributions from insurance economics, such as the work of Einav and Levin (2014), we broaden our analysis by incorporating research in other disciplines, including ethics (Steinberg 2022), law (Siegelman 2014), and medicine (Ho et al. 2020). These diverse perspectives help to provide a holistic understanding of the multifaceted implications of big data, risk classification, and privacy in insurance markets. The paper also identifies potential areas for future research, highlighting the importance of interdisciplinary collaboration between law, ethics, medicine, etc. with economics.

Traditionally, the information advantage in insurance markets resided with insured individuals, leading to the phenomenon of adverse selection. However, recent research by Brunnermeier et al. (2022) suggests that the use of advanced data analytics allows insurers to infer statistical information, effectively reversing the information advantage and the dynamics of adverse selection. Motivated by this insight, we provide an application that focuses on risk classification from the perspective of insurance companies, rather than adopting a general equilibrium model. An insurer's risk classification methodology and its accuracy can be improved through innovation in insurance pricing (Cather 2018). However, this process often requires large amounts of policyholder data and the permission to make use of such data. In addition to transaction costs that may arise from price innovation techniques (e.g., for data collection, storage and processing), insurers should consider that individuals may have different privacy preferences and potential policyholders may require some form of compensation for providing and allowing the use of their personal data (Regner and Riener 2017; Benndorf and Normann 2018; Gemmo et al. 2020). Increased privacy awareness and stricter regulation in many countries allow individuals to demand such compensation, and the application of innovations in insurance pricing can lead to changes in the customer base faced by insurers (Altman et al. 1998; Cather 2018; Lai et al. 2021).

We investigate the conditions under which insurers are willing to use policyholders' private data to classify risks more accurately. We develop a model that allows insurers to assess the maximum shift in demand for which they have a profit incentive to innovate in risk classification using private data. In doing so, we consider the "cost of privacy," which plays a central role in modern insurance markets (for a review, see Hoy 2006, and Gemmo et al. 2019). We analyze how the choice of a more accurate classification method affects the decision on the optimal number of risk classes, given heterogeneous risks in the population, and provide examples for term life insurance contracts.

The remainder of this article is structured as follows. We begin by presenting our comprehensive review of the literature in Sect. 2. In Sect. 3, we continue with an application of the use of big data in insurers' risk classification. We examine how the choice of screening technology interacts with the choice of the optimal number of risk classes using examples from term life insurance. Section 4 discusses main findings and provides an outlook for future research.

2 Literature review

We conduct a comprehensive literature review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, see Page et al. 2021) protocol to identify and categorize academic research on the use of big data in the insurance sector. The review strategy and data collection are described in Appendix 1. Based on this process, a database of 104 papers is created and key findings are extracted. The intersection of economics, business, law, ethics, and medicine has produced a rich body of literature exploring various aspects of insurance and risk management. We group all papers in four distinct areas (Table 1)[1] and shed light on the evolving landscape of insurance in the digital age, with a focus on economics and its intersections with other disciplines. To have a better understanding of the parts of research for which we have empirical results, we also add in Table 1 which papers are theoretical, empirical, and experimental.

2.1 New risks and new products

The introduction of new technologies in insurance markets has a significant impact on the frequency and severity of losses, resulting in a shift from low-severity–high-frequency to high-severity–low-frequency risks; an example is the potential tampering of self-driving cars (Eling and Lehmann 2018).[2] This transformation is driven by advancements such as automation, artificial intelligence, and interconnected systems.

[1] Other potential areas are the reduction of transaction costs (search, replication, transport, tracking, verification costs, see Goldfarb and Tucker (2019), changes in the process landscape (automated decision-making, higher efficiency, see for example, Fritzsch et al. 2021) and changes in industrial organization (economies of scale, potential disintermediation, see Eling et al. 2022). We do not study those insurance operations areas in detail to confine the focus of the paper (which centers around the intersection of big data, risk classification, and privacy). Note that we identified 104 papers, but 108 papers are listed in Table 1, because four papers (Braun et al. (2023; Brunnermeier et al. 2022; Eling and Kraft 2020;, Filipova-Neumann and Welzel 2010) are mapped into two categories. In the following discussion, we also pick up some reference from finance (e.g., Farboodi et al. 2022; Fuster et al. 2019, 2022), some classical insurance papers (e.g., Rothschild and Stiglitz 1978; Hoy 1982, 1984; Doherty and Posey 1998), and some related insurance papers on other topics (e.g., Schubert et al. 1999; Hartog et al. 2002 on risk aversion), which are outside of the 104 papers identified in Fig. 4 in Appendix 1 and thus not mentioned in Table 1.

[2] With self-driving cars, many people expect a reduction in claims number but if the driving software is hacked there could be many accidents occurring simultaneously. To further illustrate the impact of technological advancements on insurance dynamics, consider the growing prevalence of cyber threats in the digital age. With businesses and individuals relying heavily on digital platforms and interconnected systems, the insurance landscape now faces the challenge of addressing high-severity cyber-attacks that may occur infrequently but have the potential for substantial financial and operational consequences.

Table 1 Mapping of the literature

Topics	References	Key aspects
New risks and new products	Albrecher et al. (2019), Bednarz and Manwaring (2022), Biener et al. (2015)*, Bodin et al. (2018), Braun et al. (2023), Castillo et al. (2016), Cesarini et al. (2021), Cevolini and Esposito (2020), Charpentier et al. (2022), Ciborra (2006), Doss and Narasimhan (2021)*, Eling and Lehmann (2018), Faure and Li (2020), Garven (2002), Infantino (2022), Krippner and Hirschman (2022), Lanfranchi and Grassi (2022), Lindholm et al. (2022), McFall (2019), Nayak et al. (2019a), Timms et al. (2022), Xie et al. (2019)*	– Risk shift from high-frequency and low-severity to high-severity and low-frequency (Eling and Lehmann 2018) – More personalized coverage (Braun et al. 2023); exploring emerging technologies like blockchain to enhance insurability (Faure and Li 2020) – Reputational risk, driven by concerns on discrimination and negative public backlash (Fuster et al. 2019); datafication of processes with excessive data collection, posing risks to consumers (discrimination, exclusion, unaffordability) (Bednarz and Manwaring 2022)
Better/more information on insure behavior	Baecke and Bocca (2017)*, Balasubramanian et al. (2018), Barigozzi and Henriet (2011), Barry and Charpentier (2020), Bélisle-Pipon et al. (2019), Bohnert et al. (2019)*, Bologa et al. (2013), Brunnermeier et al. (2022), Che et al. (2022)*, Crainich (2017), Einav et al. (2016)*, Eling and Kraft (2020), Filipova (2006), Filipova-Neumann and Hoy (2014), Filipova-Neumann and Welzel (2010), Francois and Voldoire (2022), Geyer et al. (2020)*, Hassani et al. (2020), Hoel et al. (2006), Holzapfel et al. (2023), Hoy and Durnin (2012), Hoy and Polborn (2000), Hoy and Ruse (2005), Hoy and Witt (2007), Jin and Vasserman (2021), Keller and Transchel (2016), Leverty and Liu (2019)*, Li (2021), Li and Peter (2021), Liukko (2010), Montanera et al. (2022), McFall et al. (2020), Meyers and van Hoyweghen (2018), Nayak et al. (2019b)*, Nill et al. (2019), Paefgen et al. (2013), Peter et al. (2017), Posey and Thistle (2021), Rothstein (2015), Rumson and Hallett (2019), Saldamli et al. (2020), Thiery and van Schoubroeck (2006)	– Big data and data analytics have improved risk assessment accuracy by incorporating more data and new variables (e.g., telematics; Che et al. 2022), benefiting both insurers and policyholders (Baecke and Bocca 2017) – AI and machine learning reveal hidden patterns and relationships within large data sets, providing new insights for risk classification (Brunnermeier et al. 2022); they also enhance the detection of insurance fraud, contributing to improved efficiency (Bologa et al. 2013; Saldamli et al. 2020) – Technology addresses moral hazard through almost perfect screening mechanisms, such as telematics-based systems, incentivizing low-risk behavior (Meyers and van Hoyweghen 2018; Holzapfel et al. 2023)

Table 1 (continued)

Topics	References	Key aspects
Better risk (type) information	Barry (2020), Blasimme et al. (2019), Braun et al. (2023), Browne and Kamiya (2012), Brunnermeier et al. (2022), Cather (2018), Eling et al. (2022), Eling and Kraft (2020), Fang et al. (2020)*, Filipova-Neumann and Welzel (2010), Gidaris (2019), Guillen et al. (2019)*, Jeanningros and McFall (2020), Kiviat (2019), Krippner (2023), Liu (2023)*, McFall and Moor (2018), Palmer (2006), Południak-Gierz and Tereszkiewicz (2023), She et al. (2022), Soyer (2022), Steinberg (2022)	– Utilization of big data and advanced technology has transformed adverse selection dynamics, allowing insurers to infer statistical information and reverse information advantages (Brunnermeier et al. 2022) – On-demand insurance contracts enable better risk screening and outward shifts in utility, benefiting both insurers and policyholders (Braun et al. 2023) – Telematics and other tracking technologies offer new ways to assess risk types in property, casualty, and health insurance (Guillen et al. 2019; Eling and Kraft 2020)
Privacy, ethical concerns and legal challenges	Acquisti et al. (2016)*, Bansal et al. (2010)**, Benndorf and Normann (2018)**, Biener et al. (2020)**, Blakesley and Yallop (2019), Farrell (2012), Gemmo et al. (2019), Gemmo et al. (2020)**, Geyer et al. (2020)*, Kehr et al. (2015)**, Kim et al. (2017)*, Loi et al. (2022), Lünich and Starke (2021)*, Milne et al. (2004)*, Pew Research Center (2014), Phelps et al. (2000)*, Rohm and Milne (2004)*, Struminskaya et al. (2020)*, Tanninen (2020), Tanninen et al. (2022), Wiegard and Breitner (2019)*	– Privacy concerns play a significant role in consumers' willingness to share personal data with insurers (Phelps et al. 2000; Benndorf and Normann 2018) – Ethical considerations regarding data usage, transparency, and fairness have gained importance in the context of insurance and technology (Breidbach and Maglio 2020; Ciborra 2006) – The evolving risk landscape and increased data availability have led to ethical and regulatory discussions about privacy protection and consumer welfare (Loi et al. 2022; Gemmo et al. 2019)

* and ** denote empirical or experimental papers, respectively; papers without * or ** are theoretical or qualitative. Papers presenting simulations calibrated with empirical data are not considered empirical papers

The increasing connectivity and interdependence of systems, especially in supply chains, coupled with the collaboration of policyholders facilitated by social networks (Albrecher et al. 2019), have introduced new risks, including cyber risk. All these developments highlight the evolving nature of risks in the digital era (Eling and Lehmann 2018; Lanfranchi and Grassi 2022). The utilization of new technologies also enables insurers to offer more personalized coverage and thereby extend the insurability of risks, particularly in the case of on-demand insurance (Braun et al. 2023). For instance, the use of big data in index insurance has the potential to facilitate the development of more effective and sustainable agricultural risk management plans (Castillo et al. 2016). Similarly, big data can be used in weather index insurance (Cesarini et al. 2021) and insurance against natural disasters (Timms et al. 2022; Charpentier et al.

2022). New statistical methods can also produce different valuations of financial data according to different characteristics of investors (Farboodi et al. 2022). Other studies discuss the application of emerging technologies, such as blockchain, to approve the insurability of liability insurance in the context of 3D printing (Faure and Li 2020). The application of new technologies can also reduce barriers for consumers to enter the insurance market (Garven 2002), thereby accelerating social inclusion (Nayak et al. 2019a). Infantino (2022) provides an assessment of big data analytics from an European perspective, highlighting in particular legal and regulatory aspects.

An emerging concern in financial and insurance markets is the growing importance of reputational risk. This comprises concerns about unconscious discrimination, price discrimination, and the potential for negative public backlash, often referred to as "shit storms" (Fuster et al. 2019, 2022). The impact of machine learning algorithms on credit markets and mortgage lending underscores the need for careful management of reputational risk. There have been research efforts to eliminate potential discrimination in insurance pricing (Lindholm et al. 2022). While there is a significant amount of research underway studying the use of insurance for cyber security risk mitigation (Biener et al. 2015; Bodin et al. 2018; Xie et al. 2019; Doss and Narasimhan 2021), there is limited exploration of how the insurance sector responds to cyber risks and the ensuing reputational risk. Bednarz and Manwaring (2022) have highlighted that the datafication of insurers' processes can contribute to excessive data collection in the context of insurance contracts, with significant risks of consumer harm, particularly in terms of discrimination, exclusion, and unaffordability of insurance. Unconscious discrimination can potentially disrupt traditional characteristics of distribution and solidarity, for example, in health insurance (McFall 2019). Also sociological research examines the pricing of risks and the politics of classification in insurance and credit markets, highlighting the importance of reputation management within the insurance industry (Krippner and Hirschman 2022).

2.2 Better/more information on policyholder behavior

The use of big data and technology in insurance markets has a significant impact on the information landscape. The application of data analytics and data mining in various domains has improved the ability of insurance companies to accurately price policies (Bohnert et al. 2019; Hassani et al. 2020). This improved accuracy in risk classification is due to an increased number of observations and the inclusion of new variables in the analysis (Che et al. 2022). For example, in car insurance, the use of telematic boxes makes it possible to measure acceleration and braking behavior, which can be correlated with the likelihood of accidents. AI and machine learning techniques can reveal hidden patterns and relationships within large data sets, identifying new variables relevant to risk classification (Brunnermeier et al. 2022). The use of technology to collect data can be used to uncover risk determinants and make self-protection more effective (Li and Peter 2021). Baecke and Bocca (2017) claim that including telematic variables significantly improves the accuracy of policyholders' risk assessment. While Geyer et al. (2020) find private information to more strongly affect the bonus-malus division, they find no evidence of it affecting the

policyholders' ex ante choice contract. Brunnermeier et al. (2022) also discuss the dangers of market concentration posed by the emergence of big data (including the rise of data brokers), emphasizing the importance of consumer activism and regulatory tolerance. As noted by McFall et al. (2020), the adoption of big data analytics in insurance is transforming how risk is governed, managed, and priced within the industry. Eling and Kraft (2020) provide an extended review of the literature on the use of telematics in insurance and discuss its impact on insurability.[3]

In life, health, and long-term care insurance, the information that could be used to categorize risk includes medical tests, medical history, etc. which are considered especially sensitive. It is held that insurance discourages (prospective) policyholders from taking diagnostic tests as these tests might reveal information that leads to uninsurability (Doherty and Posey 1998). Doherty and Posey (1998) show that when linked to a treatment option, testing is encouraged when both test results and information status are restricted. In this context, changes in risk classification give rise to discussions about ethical and legal limits on the use of data, such as the debates on genetic testing or the unisex debate. (Hoy and Polborn 2000; Thiery and van Schoubroeck 2006; Liukko 2010; Rothstein 2015; Bélisle-Pipon et al. 2019; Nill et al. 2019; Posey and Thistle 2021). For instance, Hoy and Ruse (2005) emphasize that the debate over whether insurance companies should be allowed to use genetic test results for underwriting purposes must be seen in the broader context of the genetic testing debate.

A sizable part of the literature on risk classification focus on the effects of risk categorization on welfare. Hoy (1982) presents the implications of incorrectly categorizing risk on welfare. Hoy (1984) shows that categorization might lead to an increase in wealth inequality, while it reduces (on average) the unfavorable price discrimination against low risk. In addition, Hoy (2006) discusses the effect of a regulatory framework that restricts the use of certain information by insurers in rate making. He derives conditions under which regulation is explicitly welfare-enhancing or welfare-detrimental. Filipova (2006, 2007), and Filipova-Neumann and Welzel (2010) study the welfare effect of introducing insurance contracts that involve the possibility of some form of tracking data access and argue that some degree of monitoring could increase welfare. Rothschild (2011) and Dionne and Rothschild (2014) emphasize that bans on using certain information to categorize risk are sub-optimal and that alternative insurance contracts should be considered. Crocker and Zhu (2021) and

[3] The introduction of technologies such as tracking devices also has the potential to reassess risk types, not only in property and casualty insurance (e.g., driving behavior), but also in health insurance, where risk factors can be influenced by changing habits. However, the lasting effects of these changes are still subject to debate (Barry and Charpentier 2020; Meyers and van Hoyweghen 2020; Francois and Voldoire 2022). Liu (2022) quantitatively studies how the demand for artificial intelligence based on big data affects the insurance agent intermediary market. The study finds that AI demand predictions based on big data may facilitate cherry-picking for agents but fail to achieve lemon-dropping for insurers. Overall, the role of artificial intelligence is thus still limited. While more accurate risk classification does not encourage the creation of more granular risk classes per se, use of new tracking technologies can reduce the costs associated with establishing and maintaining additional risk classes (Leverty and Liu 2019; Nayak et al. 2019b; Montanera et al. 2022). The impact of digitalization on the reduction of costs of establishing and maintaining additional risk classes can in turn translate into more granular risk classes.

Pram (2021) find that utilizing a voluntary imperfectly informative test to classify risks is more efficient than not utilizing the test or making it compulsory. This result is based on the assumption that (prospective) policyholders do not know the outcome of the test ex ante. Jin and Vasserman (2021) present empirical evidence of both self-selection into monitoring and behavioral change in car insurance. They argue that monitoring generates large profits and welfare gains, but that demand frictions and policies restricting firms' ownership of collected data erode these gains.[4]

Technological advancements also offer the possibility of addressing moral hazard by implementing almost perfect screening mechanisms (Jin and Vasserman 2021; Holzapfel et al. 2023), such as telematics-based systems (see Paefgen et al. 2013; Keller and Transchel 2016; Balasubramanian et al. 2018). The integration of behavior-based personalized insurance can serve as incentives for policyholders to engage in "low-risk behavior" (Meyers and van Hoyweghen 2018). Furthermore, the application of AI and machine learning algorithms can significantly enhance the detection of insurance fraud (Bologa et al. 2013; Saldamli et al. 2020). These advancements underscore the transformative potential of technology in mitigating moral hazard and improving the efficiency and effectiveness of the insurance industry. Einav et al. (2016) highlight the economic content of risk scores, providing insights into the implications of risk assessment models on insurance markets. They find that risk scores confound underlying health and endogenous expenditure responses to insurance; even when individuals have different behavioral responses to contracts, strategic motivations for cream-skimming can persist in situations with "perfect" risk scoring within a given contract.

2.3 Better risk (type) information

Another strand of literature in the field of risk classification studies its implication on information asymmetry and adverse selection. On the one hand, Bond and Crocker (1991) argue that using endogenous categorization—classifying risks based on voluntary consumption of products that are related to the underlying loss—leads to a more efficient allocation by partly mitigating information asymmetries. Crocker and Snow (2000) add on the topic by highlighting the costs of classification risk, which depend on whether insurance markets with symmetric or asymmetric information are considered. On the other hand, Thomas (2007) emphasizes the negative effects of risk classification and argues for a socially optimal level of adverse

[4] So far, the literature has not come to a clear consensus with respect to the welfare consequences of using genetic testing in insurance pricing. On the one hand, many argue that allowing for the use of genetic testing can lead to higher social welfare (Hoel et al. 2006; Barigozzi and Henriet 2011; Peter et al. 2017; Posey and Thistle 2021). On the other hand, Hoy and Witt (2007) find only modest adverse selection costs when regulatory bans are in place, and Hoy and Durnin (2012) suggest that no significant costs from regulatory bans can be foreseen in the near future. Filipova-Neumann and Hoy (2014) explore moral hazard issues that arise in either regulatory regime, whereas Crainich (2017) examines the impact of genetic testing on self-insurance. While genetic testing can be seen as a sophisticated classification technique that an insurer can use, it is not the only one. Some methods of calculating biological age do not rely on genetic information (see for example Insilico Medicine 2023).

selection. Cather (2018) shows that innovation in risk classification methods leads to cream-skimming and pushes other insurers in the market to adopt them at a very fast pace. Browne and Kamiya (2012) study the demand for underwriting and how the cost and accuracy of categorizing tests affect it.

The utilization of big data and advanced technology in the insurance market has not only impacted risk classification, but also changed the concept of adverse selection, resulting in reverse selection dynamics (Filipova-Neumann and Welzel 2010; Cather 2018; Eling et al. 2022). As early as 1976, Rothschild and Stiglitz showed that one way to deal with adverse selection is to distinguish high-risk and low-risk individuals, thereby establishing a separating equilibrium. Brunnermeier et al. (2022) point out that insurance companies transfer information advantages from the insured to the insurance company by inferring statistical information, that is, the reversal of adverse selection.[5] Braun et al. (2023) show that the heterogeneity of policyholders in terms of claim amounts and claim frequency can be better exploited through on-demand contracts, which allow for better screening of the policyholder type. Furthermore, telematics can be beneficial for high-risk individuals as a condition of insurability, effectively mitigating selection problems (Guillen et al. 2019; Eling and Kraft 2020; Fang et al. 2020; She et al. 2022). Jeanningros and McFall (2020) investigate the value of sharing data in a life and health insurance company, also highlighting the role of branding and behavior in insurance markets. These transformations in insurance markets have given rise to ethical and legal considerations regarding data usage, contributing to the ongoing discourse on the topic (Palmer 2006; Kiviat 2019; Steinberg 2022; Krippner 2023; Południak-Gierz and Tereszkiewicz 2023). For instance, concerns have been raised about the potential overuse of medical data by insurance companies and its potential impact on medical advancements (Blasimme et al. 2019). Some studies argue that insurance companies collect customer data through wearable devices and other means, resulting in consumers relinquishing power and control over the data generated from their activities (Gidaris 2019). The ethical implications of data-driven business models are also examined by Breidbach and Maglio (2020), who analyze accountable algorithms and the ethical considerations associated with their use. They highlight the need for transparency and fairness in algorithmic decision-making. Similarly, Ciborra (2006) explores the ethical dimensions of risk and digital technologies, highlighting that digital tools are both the infrastructure of the risk industry and the source of new, often unpredictable, risks. In particular, Liu (2023) found that AI-generated demand information reduces sales agents' own information acquisition and increases adverse selection; agents using AI attract riskier consumers and do not match them to more expensive products to achieve stronger incentive compatibility.

[5] There are earlier theoretical papers by Villeneuve (2000, 2005) on the insurer knowing more than the policyholder. Both papers discuss the consequences of insurance firms evaluating risk better than customers, in a monopolistic Villeneuve (2000) and in a competitive Villeneuve (2005) situation.

2.4 Privacy, ethical concerns, and legal challenges

The utilization of new technologies has instigated shifts in risk perception and raised concerns regarding privacy and transparency within insurance markets (Gemmo et al. 2019). Several authors have tried to identify the characteristics that influence consumers' willingness to share personal data, be it the type of information, the characteristics of the company collecting and using the data, the purpose of use, or the consumers' own characteristics (Phelps et al. 2000; Rohm and Milne 2004; Milne et al. 2004; Pew Research Center 2014; Acquisti et al. 2016; Benndorf and Normann 2018). Farrell (2012) proposes a model that regards privacy as a final good whose optimal level can be chosen efficiently, whereas Kehr et al. (2015) suggest that behavioral biases affect the privacy valuation. The literature empirically observes differences between the willingness to sell private data and the willingness to buy privacy protection (Phelps et al. 2000; Milne et al. 2004). Additional studies explore issues that may arise from the application of big data in the insurance industry, such as the transformation of fairness connotations (Barry 2020), the dynamics between individuals and groups (McFall and Moor 2018), and the enhancement of privacy protection laws for consumers (Soyer 2022), among others. Related to this, Strohmenger and Wambach (2000) and Hoy and Ruse (2005) provide a discussion of the arguments for and against the use of genetic testing in insurance rating. Some studies argue that behavior-based insurance can exploit the insured (Tanninen 2020) and potentially compromise the autonomy of policyholders (Tanninen et al. 2022).

Some papers study the willingness to share data with insurance firms specifically.[6] Wiegard and Breitner (2019) consider wearable technologies and suggest that privacy concerns are the main hurdle for pay-as-you-live insurance contract adaptation. Blakesley and Yallop (2019) conduct a similar study on the UK insurance market. They emphasize that insurance firms should establish ethical standards above the legal requirements for data-driven insurance contracts to achieve wider consumer adoption against a 'fair' incentive. Gemmo et al. (2019) consider an insurance market framework with asymmetric information, and show how the existence of policyholders' privacy concerns can affect market equilibria and social welfare. The authors find that information disclosure can lead to a Pareto improvement of social welfare—even in the presence of privacy costs—although it can also decrease or eliminate cross-subsidies. Striking a balance between the desire for privacy and

[6] Empirical research suggests that insurers typically face a decreasing demand function (Einav et al. 2010). Risk aversion is one possible explanation for differences in willingness to pay between (potential) policyholders with similar risk characteristics. A large body of research attempts to identify consumer characteristics that are associated with higher risk aversion, such as gender, age, etc. (Schubert et al. 1999; Hartog et al. 2002). Some of the consumer characteristics empirically associated with risk aversion are also shown to be linked with the willingness to share data (Bansal et al. 2010; Kim et al. 2017; Benndorf and Normann 2018; Geyer et al. 2020; Struminskaya et al. 2020; Gemmo et al. 2020). The literature studying the relationship between risk aversion and willingness to share data is limited and its findings are non-conclusive. Gemmo et al. (2020) do not find a significant association between general risk aversion and the probability of agreeing to sell private data, whereas Biener et al. (2020) find indication that higher risk aversion is associated with a lower willingness to pay for policies that require sharing additional data.

the necessity to mitigate risks presents challenges for individuals and the industry alike (Biener et al. 2020). The evolving risk landscape and the increased availability of data can impact the market structure (Gemmo et al. 2019), while also give rise not only to ethical, but also regulatory considerations (Blakesley and Yallop 2020; Loi et al. 2022). This also raises the question of how to reconcile consumers' perceived privacy risks with their own welfare, which is influenced by various factors (Wiegard and Breitner 2019; Lünich and Starke 2021).

All these developments in the risk landscape underscore the profound impact of technology on the insurance industry, requiring careful consideration of risk management strategies and the establishment of regulatory frameworks to effectively address the challenges and ensure the existence of fair and sustainable insurance markets. The inverse selection dynamics, that is, the transfer of information advantages from the insured to the insurer, opens a broad area of future research that revisits its results from (standard and non-standard) models with asymmetric information, in which the informational advantage has been on the side of the policyholder. In the subsequent section, we provide an example of a framework that considers the firm's perspective in deciding whether and to what extent to implement new data-driven technologies. Our application analyzes the decision of an insurance company to choose the risk classification system that maximizes its expected profit. We consider the privacy implications of using private data in the insurer's decision-making process by relating the willingness of (potential) policyholders to share private information to their willingness to pay for insurance. With this we connect two fundamental parts of the literature review; more accurate risk classification (third part) and reduction in demand from privacy concerns (fourth part).

3 Application: the optimal risk classification system from the insurer's perspective

In Gatzert et al. (2012), different forms of substandard annuities are presented and the challenges of the underwriting process in insurance practice are identified. In a theoretical model, a risk classification system for substandard annuities is derived assuming that the insurer wants to maximize it's expected underwriting profits and that risk classification is costly. In addition, the model includes the cost of an inappropriate risk assessment (causing underwriting risk) that occurs when policyholders are assigned to inappropriate risk classes. Specifically, such inadequate risk assessment is modeled by assuming error probabilities for misclassifying policyholders into a lower risk class, thereby understating expected indemnity payments.

We aim to contribute to the existing body of knowledge by linking the willingness of (prospective) policyholders to provide private information for risk classification purposes to their willingness to pay for insurance. We combine the analysis of an insurer's optimal risk classification strategy with considerations of policyholders' privacy preferences. Given the developments toward better risk predictability—although the debate on the welfare effect and the best regulatory framework is still ongoing—we find it interesting to examine the decision

to implement new technologies for risk classification purposes from the firm's perspective. We add to this area of research by analyzing an insurer's underwriting decision process with respect to offering policies that require policyholders to share private data in exchange for some compensation. The additional data could allow insurers to classify risks more accurately.

In the absence of classification costs and under full information, it would be optimal for the insurer to classify each subpopulation of policyholders with equal risk into a separate group (Gatzert et al. 2012). In practice, the classification process involves transaction costs, and the information available to insurers to identify which risk subpopulation a (prospective) policyholder belongs to is not perfect. Therefore, a decision must be made regarding the optimal classification system.

We analyze the choices an insurer faces when implementing new screening techniques that can improve risk classification. To that end, we take the position of an insurance company that seeks to maximize its expected underwriting profit. We provide a framework that an insurer could use to decide whether and to what extent they should invest in pricing innovation using new technologies and big data analyses. Moreover, we revisit the problem of choosing the optimal risk classification system presented in Gatzert et al. (2012) and extend it to incorporate the conditions under which innovation in risk classification methods is profitable and how the optimal classification system changes with it.

We build an application for term life insurance business. Typically, the insurer could group policyholders into different risk classes based on estimates of their mortality risk with a certain classification error. Based on the heterogeneity of the underlying population and their price-demand characteristics, the insurer would choose to offer a profit-maximizing number of classes (Gatzert et al. 2012). Studies suggest that biological age[7] serves as a good predictor of age-related diseases and mortality risk (Horvath 2013; Putin et al. 2016; Huang et al. 2017; Milevsky 2020a; Wu et al. 2021). Therefore, requesting policyholders' data necessary to calculate their biological age can lead to improved accuracy in the classification into risk classes. However, the requirement to share personal data is expected to affect the price–demand characteristics, because the updated policy embeds both term life coverage and trading personal data. This way, the decision of the insurer on whether to use risk class indicators, such as biological age, affects their decision on the optimal number of risk classes to offer. In the following sections, we set up a framework for profit-maximizing insurers to navigate through these decisions.

[7] The derivation of an individual's biological age is based on the identification of "biomarkers of aging" that provide "better estimates of expected remaining lifetime and future mortality rates" (Huang et al. 2017, p. 58). Different concepts of how to approximate biological age have been laid out in Milevsky (2020b, pp. 149–151).

3.1 General procedure of selecting a classification system to maximize the expected profit

We lay out a framework to analyze the insurer's problem of selecting the classification system that maximizes its expected underwriting profit. The proposed framework focuses on two choice variables, namely, the number of risk classes offered, and the probability of misclassification. To focus on the interaction of these two choice variables, we make some simplifying assumptions:

(i) We assume that risks are purely unsystematic and hence, the owners of the insurance company can (fully) diversify them. Therefore, our choice of the "optimal" classification system refers to the classification system that maximizes the insurer's expected profit.[8]
(ii) We assume no other risk source beside the policyholders' claim distributions.
(iii) We set the riskless rate of return in our two-points-in-time-model to zero.
(iv) We assume that the insurer cannot go into default within the timeframe of our model.
(v) We assume that the insurer faces a downward-sloping, linear demand function in each of the subpopulations.
(vi) We consider no general administrative or agency costs.

Since potential policyholders within a group are homogenous in terms of risk, the actuarially fair premium (per contract) based on the expected claims is identical. This is reflected in a constant (parallel) line of marginal costs. Moreover, the insurer has a risk of misclassification—hence, the policyholder's risk type is not fully transparent. The insurer does not have the full information regarding the risk group to which the potential policyholder belongs, it can only infer potential policyholders' risk groups based on the information it is allowed to gather from them and its internal risk evaluation models. Misclassification can generate a loss (or at least a deviation from the maximal profit attainable) to the insurer if, for example, a high-risk is categorized as a low risk.[9] Furthermore, also potential policyholders are not fully aware of their risk group. Potential policyholders could have access to their data when performing given tests, but they would not have access to the data-extensive risk evaluation models that the insurer uses. Therefore, while the potential policyholder may have an indication of the risk group to which she belongs, that self-assessment will not always be correct.

It is assumed that introducing new screening techniques that require policyholders to provide additional personal data, will, on the one hand, result in a lower or equal reservation price for the new policy[10] from the potential policyholder's perspective

[8] Our setting also holds true for (partly) systematic underwriting risk if the insurer is assumed to behave risk neutral.

[9] We assume that misclassification can happen to either direction, lower or higher risk class. Appendix 2 illustrates how each direction of misclassification impacts the profit and discusses the net effect.

[10] The reservation price for the new policy is the difference between the reservation price for insurance coverage and the (non-negative) reservation price for sharing private data.

(Regner and Riener 2017; Benndorf and Normann 2018; Gemmo et al. 2020). On the other hand, new screening techniques are expected to improve the accuracy of the classification system, giving the insurer the possibility to identify with a better accuracy the risk group to which the potential policyholders belong (Baecke and Bocca 2017; Verbelen et al. 2018; Geyer et al. 2020). In addition, an insurer faces the possibility of losing or acquiring policyholders to/from competitors who do not offer the product with the same accuracy in risk classification.

We work through the problem of analyzing the trade-off between the effects of the new classification technique by first constraining the expected impact of the private data requirement on the demand curve. Second, we describe the general decision algorithm of an insurer faced with an underlying population composed of n risk groups and, choosing the classification system that maximizes its expected profit. In the decision algorithm, we analyze how the choice of screening technology interacts with the choice of the number of risk classes to offer. In addition, an illustrative numerical application using data from the German term life insurance market is provided.

The choice of the classification system refers to the simultaneous choice of the number of risk classes and the classification method, where the latter may or may not require the use of private data. The choice of whether to require the use of private data determines the accuracy of the classification system. The insurer is constrained in this choice to the extent that, with full accuracy, a given (estimated) downward shift in the demand curve is expected to occur. The demand curve for the insurance policy that requires the use of private data and guarantees full accuracy in classification differs from the demand for the initial policy for two main reasons. First, policyholders require compensation for the additional personal information that they need to provide. Thereby, their willingness to pay for the new policy changes. To determine how this change reflects in a new demand curve, we need to consider the relationship between willingness to pay for insurance and privacy concerns (translated into a required deduction). While there is, to the best of our knowledge, no conclusive research on this relationship, the fact that they are both influenced by a very similar set of consumer characteristics suggests that the two might not be independent (Bansal et al. 2010). We assume the willingness to pay for insurance to be negatively related to the willingness to share private data and to grant permission for their use in the risk classification process.[11] In this case, the demand curve will shift downwards by more for those who have a higher willingness to pay for insurance. One way to interpret this would be to consider willingness to pay for insurance as driven by the degree of risk aversion[12] and thinking of more risk-averse policyholders as more prone to assessing private data as sensitive. This would lead

[11] In the following, when using the term "sharing private data" we imply that sharing involves granting permission to use the data for the purpose of risk classification.

[12] This holds if we assume that all policyholders have the same preference function.

policyholders with a higher degree of risk aversion to request a higher deduction to compensate for the disutility[13] caused by sharing the required private data.[14]

Second, the change in the risk classification methodology would have an impact on the insurers' competitiveness in the market, provided that the insurer has proprietary rights over the new methodology and the additional data collected. This will result in the loss of some potential clients and the acquisition of others. The magnitude of this effect depends on the timing of implementation of the new risk classification technology versus competitors. If the insurer is among the first movers, on the one hand, the incentive is for potential policyholders who assess themselves as belonging to the lower risk groups to switch from competitors to the insurer applying the new classification system. On the other hand, potential policyholders who assess themselves to belong to the higher risk groups have the incentive to switch to other "traditional" providers. Literature on adverse retention suggests that low-risk policyholders are more likely to switch providers (Altman et al. 1998; Lai et al. 2021). However, there is also evidence suggesting that the first-mover advantage is minor (Reimers and Shiller 2018) and therefore it is fair to assume that either type of shift is limited. Whereas, if the more accurate risk classification methodology has already been implemented by many competitors, the expected effect is more on retaining policyholders from lower risk groups.

We assume that within a risk group, there is no interrelation between the willingness to pay for the new insurance policy and the predisposition to change providers. To the best of our knowledge, there is no research indicating otherwise. To pin down the shape of the demand shift, we assume that policyholders with a willingness to pay for insurance equal to zero require no compensation for sharing private data. Therefore, the willingness to pay for the new product does not become negative. Figure 1 depicts, in a given risk group, the demand for the new insurance product that uses private data to achieve fully accurate risk classification, versus the demand for the insurance product that uses a conventional risk classification methodology that does not require private data. We denote N_s the total number of policyholders with a positive willingness to pay for insurance in risk subpopulation s, P_s^{R} the maximal reservation price of policyholders in risk subpopulation s, and P_s^{A} the actuarially fair premium for policyholders in risk subpopulation s. We denote by $1-\alpha_1$ the percentage discount that the policyholder with the highest willingness to pay for insurance requires, leading to a new highest willingness to pay of $\alpha_1 P_s^{\mathrm{R}}$. As the willingness to pay for insurance and the discount required for giving up private data are positively related, the new demand curve will be obtained by multiplying the initial one with a coefficient $0 < \alpha_1 < 1$. The dark blue line in Fig. 1 depicts the shift in demand driven solely by the change in willingness to pay for insurance of the initial policyholder base, that is, if no client migration was expected. The parallel shifts in demand driven by the effect on market competitiveness is described by the

[13] We model the disutility caused by sharing private data via a negative cash flow on the utility of the policyholder.

[14] Appendix 2 shows how alternative assumptions regarding the relationship between willingness to pay for insurance and willingness to share data would affect our analyses.

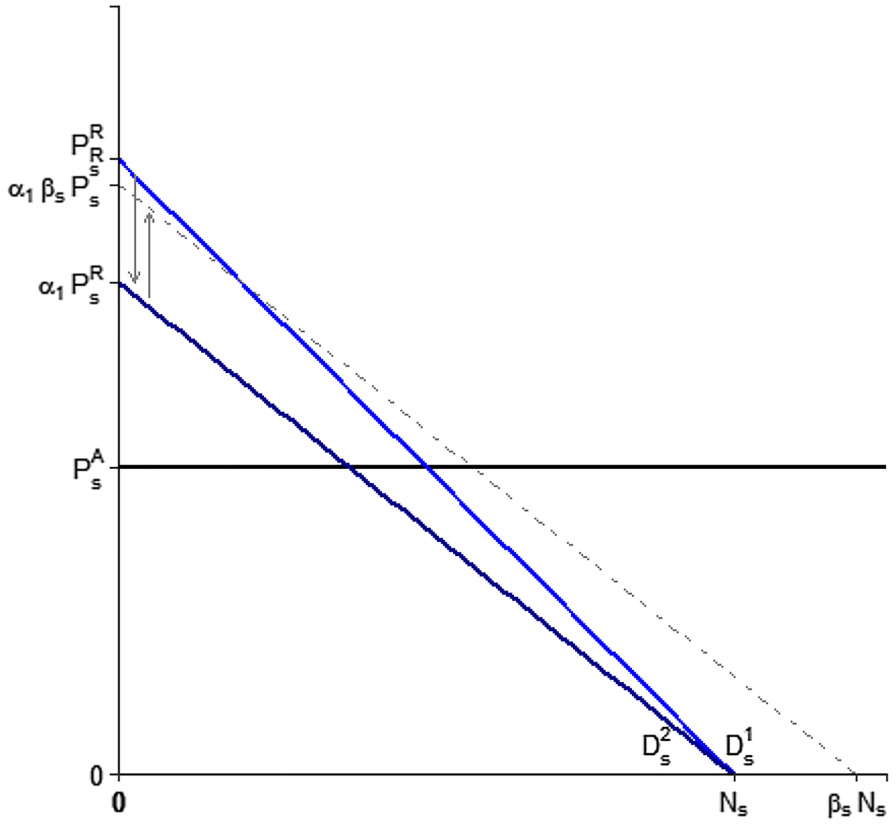


Fig. 1 Demand curve shift in a given risk group when using private data in risk classification. This figure illustrates the cumulative shift in demand in a given risk group, accounting for the change in demand driven directly by the data requirement within the client base of the insurer as well as the additional shift caused by the exchange of potential policyholders in a risk group among insurers

coefficients β_s. Note that the parallel shift depicted by the dashed line is only illustrative and, the actual shift could be an expansion or a contraction.

We build on the model set forth by Gatzert et al. (2012) by attaching a binary choice of misclassification probability to the classification system. The insurer can either rely on given information/data constraints and classify risks with a certain misclassification probability or opt for an innovative classification method that attains full classification accuracy. The latter is possible only when requesting prospective policyholders to provide and consent on the use of certain private information. As discussed, linking the insurance policy to the request for use of private data leads to an alteration of the demand—due to compensation for privacy concerns associated with sharing private data and effects on competitiveness. We lay out an optimization procedure for the insurer to choose the classification system that maximizes the expected profit and analyze how the given variables affect this decision.

We examine the incentives in terms of increased expected profit of insurers to innovate in the risk classification space.

3.2 Risk classification framework

We consider S heterogeneous subpopulations that contain N_s policyholders, with $s \in [1, S]$, who are homogeneous with respect to the expected claim payment for a certain type of insurance policy.[15] For example, in the case of term life insurance, we can imagine the overall population of prospective policyholders formed by subpopulations with the same life expectancy.[16] Subpopulations are characterized by their cost function as well as their price-demand function. Since we consider policyholders within a subpopulation homogenous in terms of expected claim payment, the respective average cost (and marginal cost) function will be constant and equal to the expected claim payment per policyholder in the subpopulation, denoted by P_s^{A}.

A classification system m is considered any grouping of all subpopulations into I_m risk classes. In this setting, when ranking risk subpopulations in decreasing order, only adjacent subpopulations can be grouped into the same risk class. Moreover, we assume that the number of subpopulation(s) per risk class is equal among risk classes, when possible, otherwise higher risk class(es) include one more subpopulation than lower risk class(es). The problem then consists of finding the optimal number of risk classes, between 1—that is, grouping all subpopulations together—and S—that is, putting each subpopulation in a separate class. Risk classes will also be characterized by their cost function and price–demand function. In the cases in which a risk class contains more than one subpopulation, its cost and price–demand function will be aggregated functions of the cost and price–demand function of the contained subpopulations.[17] In the case of S risk classes, the cost and price–demand functions of the risk class will be the same as that of the corresponding subpopulation.

A classification system m, with I_m risk classes will also be characterized by classification costs, denoted by C, and the probability of misclassification, denoted by ur. We assume classification costs to be proportional to the number of risk classes and model them as $C = c(I_m - 1)$, where $c \in R_0^+$. For simplicity, we assume that policyholders are only misclassified to adjacent risk classes and that the probability of misclassification is equal in either direction. This implies that the highest and lowest-risk class will have a lower total probability of misclassification as they only have one adjacent risk class. Furthermore, in this setting, the misclassification probability can only take two possible values: $r > 0$, when using the default classification

[15] Note that since we assume underwriting risk to be unsystematic, and we have no other risk source, the expected claim payment fully defines a risk subpopulation.

[16] Not considering special contract features, policyholders with the same life expectancy can be thought of as having the same actuarially fair premium when it comes to term life insurance.

[17] The aggregation process is complex and must be conducted stepwise. When linear price-demand functions are assumed, it can be conducted in line with what is presented in Sect. 3.2 by Gatzert et al. (2012).

methodology, or 0, when using an innovative classification methodology that employs private data. The innovative classification methodology is associated with a shift in the price–demand curve in each risk subpopulation. Hence, a classification system can be fully defined by the number of risk classes, the accuracy of classification (reflected in the potential use of private data), and classification costs. The classification system can be denoted by $m\{I_m, \mathrm{ur}(\alpha_1, \beta^S), C(I_m)\}$,[18] where β^S is a vector of length S, containing the coefficient of expansion or contraction of the client base in each risk subpopulation, for simplicity we will refer to this only as a classification system m.

Within a classification system, risk classes are characterized by their cost function and price–demand function. We will denote $\mathrm{MC}^l(n)$ the marginal cost function of risk class l, where $l = \{1, \ldots, I_m\}$, and $\mathrm{WTP}^l(n)$ its price–demand function, that is, willingness to pay function. These are functions of the number of risks in the risk class and, where the risk class contains more than one subpopulation, are obtained by aggregating the corresponding functions of the subpopulations. In what follows, we will omit the l subscript as well as the n and, for simplicity, write $l\{\mathrm{MC}, \mathrm{WTP}\}$ to refer to a risk class. Having a defined demand and cost function, each risk class will also have a profit function, which apart from the WTP and MC functions, depends also on the probability of misclassification of policyholders from that risk class into adjacent risk classes—and the price the misclassified policyholders are offered in the "incorrect" (adjacent) risk class—if the probability of misclassification is positive.

In this setup, the problem of finding the optimal, that is, the profit-maximizing classification system from the insurer's perspective can be broken down into several steps:

1. Given the classification system $m\{I_m, p(\alpha_1, \beta^S), C(I_m)\}$ with I_m risk classes, classification cost C, and a classification methodology that either makes use of private data or not, and translates into a combination of misclassification probability and demand shift, calculate the overall profit π_m based on one of the procedures below, as appropriate:

    (i) In the case of the default classification methodology that results in a misclassification probability $\mathrm{ur} = r > 0$ (and $\alpha_1 = 1$, $\beta^S = 1^S$):

    (1) Rank subpopulations by riskiness (highest to lowest) and separate all subpopulations into risk classes l, where $l = \{1, \ldots, I_m\}$ (l= 1 highest risk, $l = I_m$ lowest risk).

    (2) For each risk class l, calculate the price–demand and cost functions by aggregating the corresponding functions of the subpopulations that it contains, and find the profit-maximizing price-demand combination $p_l^{\mathrm{d}*}$ and $n_l^{\mathrm{d}*}$, hereon the superscript d refers to variables under the default classification methodology.[19]

[18] The probability of misclassification, ur, is expressed as a function of the demand shift coefficients to denote that the probability of misclassification chosen comes with a given demand shift.

[19] This set of prices maximizes expected profit disregarding anticipated misclassification and might differ from the set of prices that maximizes total profit in the presence of (anticipated) misclassification $P_l^{**} = \mathrm{argmax} \sum \pi_l(P_l)$ for $l = \{1 \ldots I_m\}$; However, the latter has no analytical solution and can only be

(3) For each risk class l, calculate the additional demand created by incorrectly offering to the policyholders belonging to it the optimal prices of the adjacent risk classes $p_{l-1}^{\mathrm{d}*}$ and/or $p_{l+1}^{\mathrm{d}*}$, $n_{l-1,l}$, and $n_{l,l+1}$, respectively.[20]

(4) Calculate the maximal expected profit for each risk class and then adjust for misclassification as $\widetilde{\pi}_l^{\mathrm{d}} = r\left(n_{l-1,l}\left(p_{l-1}^{\mathrm{d}*} - p_l^{\mathrm{A}}\right)\right) + (1-2r)\left(n_l^{\mathrm{d}*}\left(p_l^{\mathrm{d}*} - p_l^{\mathrm{A}}\right)\right) + r\left(n_{l,l+1}\left(p_{l+1}^{\mathrm{d}*} - p_l^{\mathrm{A}}\right)\right)$ $r\left(n_{l,l+1}\left(p_{l+1}^{\mathrm{d}*} - p_l^{\mathrm{A}}\right)\right)$ when the risk class has two adjacent risk classes; $\widetilde{\pi}_l^{\mathrm{d}} = r\left(n_{l,l+1}\left(p_{l+1}^{\mathrm{d}*} - p_l^{\mathrm{A}}\right)\right) + (1-r)\left(n_l^{\mathrm{d}*}\left(p_l^{\mathrm{d}*} - p_l^{\mathrm{A}}\right)\right)$ or $\widetilde{\pi}_l^{\mathrm{d}} = r\left(n_{l-1,l}\left(p_{l-1}^{\mathrm{d}*} - p_l^{\mathrm{A}}\right)\right) + (1-r)\left(n_l^{\mathrm{d}*}\left(p_l^{\mathrm{d}*} - p_l^{\mathrm{A}}\right)\right)$ for the highest and lowest-risk class, respectively; or $\widetilde{\pi}_l^{\mathrm{d}} = \pi_l^{\mathrm{d}} = n_l^{\mathrm{d}*}\left(p_l^{\mathrm{d}*} - p_l^{\mathrm{A}}\right)$ in the case of only one risk class.

(5) Calculate the total expected profit under this classification system as the sum of the expected profits in each risk class after deducting the classification costs: $\pi_m^{\mathrm{d}} = \sum_{l=1}^{I_m} \widetilde{\pi}_l^{\mathrm{d}} - c(I_m - 1)$.

(ii) In the case of the innovative classification methodology that uses private data and yields a fully accurate classification ur = 0:

(1) Estimate the shift in demand that the incorporation of data requirements in the insurance policy would cause, both in terms of affecting the willingness to pay for the policy of the current client base in terms of magnitude $0 < \alpha_1 < 1$, and the effect on client migration from/to competitors β^S.

(2) Rank subpopulation by riskiness (highest to lowest) and separate all subpopulations into risk classes l, where $l = 1, \ldots, I_m$ (l= 1 highest risk, $l = I_m$ lowest risk).

(3) For each risk class l, calculate the price–demand and cost functions by aggregating the corresponding (shifted) functions of the subpopulations that it contains and find the profit-maximizing price–demand combination $p_l^{\mathrm{n}*}$ and $n_l^{\mathrm{n}*}$, hereon the superscript n refers to variables under the innovative classification methodology.

Footnote 19 (continued)

solved by brute force trial error which becomes very inefficient in the general case of n risk classes. As we are working toward a framework that allows the choice of the profit-maximizing classification system in the general case where the underlying population can have S subpopulations, we will proceed by comparing it to the set of expected profit-maximizing prices set not considering anticipated misclassification, P_l^*. In specific applications, when the insurer has chosen a different set of prices in the presence of misclassification, $P_l^{**} \neq P_l^*$, such that $\sum \pi_l\left(P_l^{**}\right) > \sum \pi_l\left(P_l^*\right)$, one should compare the profit derived from the classification system with improved accuracy to $\sum \pi_l\left(P_l^{**}\right)$. Our results in this paper will generally differ compared to those in a case where anticipated misclassification is considered when setting the prices in the default classification system in that the maximal discount offered to (prospective) policyholders (or maximal fall in demand accepted) will be larger. However, our results can still serve as an upper bound in guiding the decision-making process.

[20] When referring to the additional demand created by misclassification, we omit the superscript d, as we assume no misclassification (therefore, no additional demand created) in the alternative methodology.

(4) For each risk class l, calculate the maximum expected profit $\pi_l^{\mathrm{n}} = n_l^{\mathrm{n}*}\left(p_l^{\mathrm{n}*} - p_l^{\mathrm{A}}\right)$.

(5) Calculate the total expected profit under this classification system as the sum of the expected profit in each risk class deducting classification costs: $\pi_m^{\mathrm{n}} = \sum_{l=1}^{I_m} \pi_l^{\mathrm{n}} - c(I_m - 1)$.

2. Repeat this procedure for all 2 * S possible classification systems $m^{\mathrm{d}} \in \left\{1^d, \ldots, S^d\right\}$ and $m^{\mathrm{n}} \in \{1^n, \ldots, S^n\}$ and choose the one that yields the highest total profit, m^*.

Following this procedure, the insurer can simultaneously decide on whether it is optimal to innovate the classification technique using private data and choose the optimal number of risk classes. Figure 2 illustrates the discussed algorithm for selecting the profit-maximizing classification system with respect to the number of risk classes and methodology employed.

The proposed procedure has its limitations, partly stemming from the simplifying assumptions we made such as linearity of the demand curve of the insurer, no risk of default of the insurer, symmetric misclassification, composition of the underlying population, etc. While these assumptions might need to be relaxed/adapted in

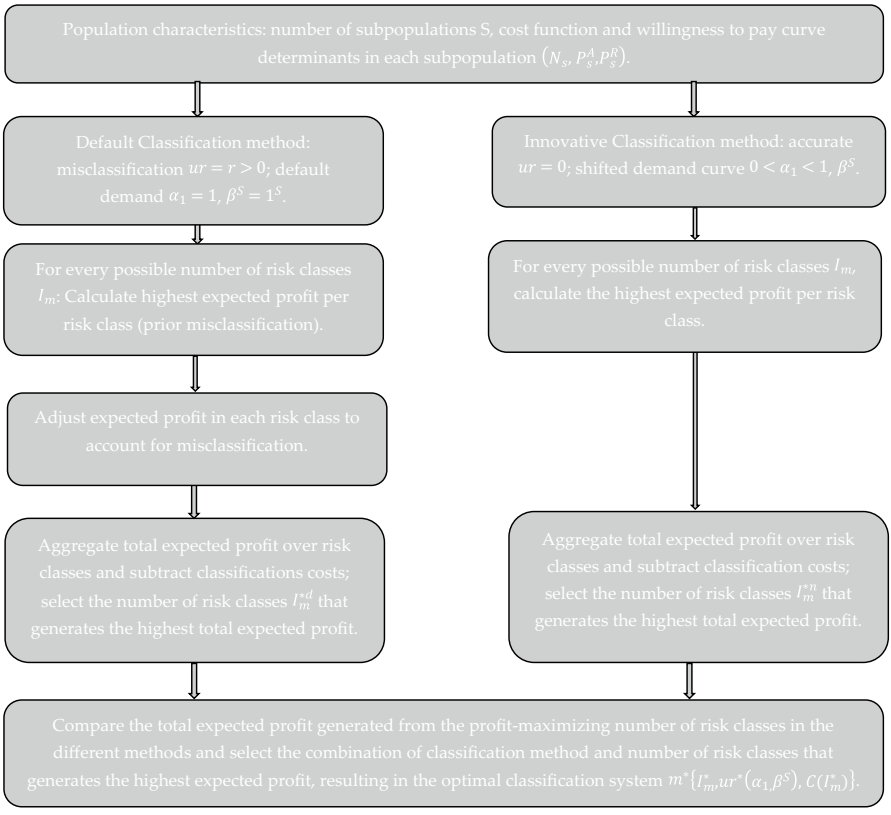


Fig. 2 Algorithm of selection of the profit-maximizing classification system

practical applications, they allow us to assess the average effects of the usage of new technologies on the profit-maximizing number of risk classes without considering specifically the steepness of the demand curve at the initial and new profit-maximizing combination in each risk class or the policyholders' willingness to change providers based on their differences in default risk.

Figure 2 illustrates the algorithm for selecting the classification system that maximizes expected profit by integrating the selection of the classification methodology, employing additional data or not, and the number of risk classes offered.

3.3 Numerical applications

We provide two illustrations of our proposed procedure for selecting the optimal risk classification system. These examples allow us to discuss more concretely the incentives for insurers to innovate in risk classification and how this affects the incentives to offer different granularities of risk classes.[21]

3.3.1 Population of five homogeneous subpopulations in the term life insurance market

The application of our proposed decision-making process requires an estimation of market data specific to the firm, the line of business under consideration, and client characteristics in that line of business. However, to illustrate our proposed procedure, we will apply our setup to the term life insurance market, with estimates taken from empirical data collected by Braun et al. (2016).

An important risk factor used to classify policyholders when it comes to term life insurance is age. Age is used to estimate mortality risk. However, another risk factor, biological age, has caught insurance researchers' interest (Hochschild 1988). Recent research shows that biological age is a better predictor of mortality than chronological age (that is, the time that has passed since a person was born) (Huang et al. 2017; Mamoshina et al. 2018; Milevsky 2020a; Wu et al. 2021). Based on this, we could assume that insurers can opt either for classification into risk classes based on (chronological) age or request policyholders to provide information and test data necessary to estimate their biological age. The former would lead to a misclassification probability in terms of predicting mortality, while the latter allows the insurer to classify policyholders more accurately into mortality brackets. For simplicity, we assume that the second method is fully accurate.

We refer to the data that Braun et al. (2016) collected on the willingness to pay for the "classic product" divided into five groups based on age. We concentrate only on policyholders who smoke, to make sure that the subpopulations are approximately homogeneous. To align with our setup, we derive the linear approximation of the willingness to pay for term life insurance curve in each subpopulation and take the highest reservation price and the number of policyholders per group from

[21] The R code used to compute the numerical application is available in the Online Appendix.

the approximation. The marginal cost in each subpopulation is taken as constant, equal to the average variable costs presented by Braun et al. (2016). Note that in the data, ranking the groups by marginal cost does not yield the same result as ranking them by the highest reservation price. To aggregate the demand curves correctly, we rank the groups based on the highest maximal reservation price. Classifying based on age leads to a 20% misclassification probability in either direction. In addition, the classification cost increases by $c = 100$ units with the number of risk classes. Moreover, we assume that the introduction of the innovative classification method is not expected to lead to an exchange of clients with competitors in either subpopulation, that is, $\beta^5 = 1^5$ and that willingness to pay for insurance is negatively correlated to the willingness to share data. Under these assumptions, the maximal profit generated in either classification system is presented in Table 5 in the Appendix. Note that in the absence of classification costs, it is optimal for the insurer to classify each subpopulation into a separate risk class, regardless of classification accuracy. Once classification costs are introduced, applying classification based on age would lead to a division into four risk classes yielding the highest profit. The insurer would be able to accept a maximal fall in demand with $\alpha_1 = 0.946$ to implement a classification based on biological age (under the assumption that classifying based on biological age eliminates misclassification) and find it at least as profitable as classification based on chronological age. In this case, the optimal number of risk classes would be two. Note that the cost of misclassification, that is, the difference between the total (optimal) profit without misclassification and the total profit with misclassification, decreases with the number of risk classes. Therefore, the increased profit from a more accurate classification is higher in classification systems with fewer risk classes.

3.3.2 A numerical example of a population with seventy homogenous subpopulations

In practice, larger variation among policyholders is common, and assuming that the population has only five subpopulations is not very realistic. Therefore, we relax the assumption imposed by the availability of willingness-to-pay data and construct an example that illustrates the problem of selecting the profit-maximizing classification system when allowing for more diversity among (prospective) policyholders in a population.

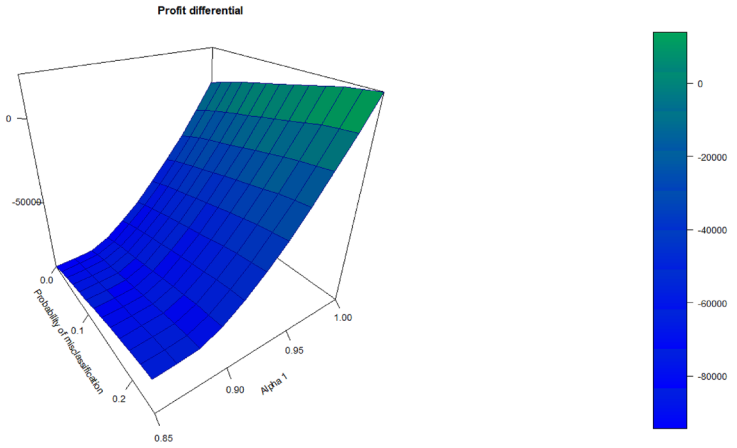
We consider a population made of 70 subpopulations ($S = 70$). This assumption regarding the diversity within the population of potential policyholders is realistic, for instance, in the case of term life insurance, where mortality risk differs every year of biological age, ceteris paribus, but within a subpopulation of the same biological age, it can be considered constant. In the absence of empirical data regarding willingness to pay at this level of granularity, we construct a numerical example. We assume a vector of maximum reservation prices per class decreasing from 3500 in the highest-risk subpopulation to 50 in the lowest-risk subpopulation, an equal number of policyholders in each subpopulation $N_s = 100$ and expected claim payments decreasing from 2800 in the highest-risk

subpopulation to 40 in the lowest-risk subpopulation.[22] For a given range of initial probabilities of misclassification, up, and a possible range of classification cost per additional risk class, c, Table 6 in the Appendix shows the number of risk classes that would yield the highest total profit in each case. Furthermore, Table 6 presents the maximum shift in demand (measured by the coefficient α_1) that the insurer is willing to accept to obtain policyholder data if the insurer estimates no effect on client migration to/from competitors in either subpopulation, that is, $\beta^{70} = 1^{70}$. Lastly, column 5 in Table 6 presents the number of risk classes that yield the highest profit given a shift in demand due to the acquisition of policyholder's data, necessary to implement a fully accurate classification. In term life insurance, we can think of an initial misclassification probability ur when using age to proxy mortality risk. Then suppose that by acquiring from policyholders the data necessary to calculate biological age, the insurer would be able to eliminate this probability of misclassification.

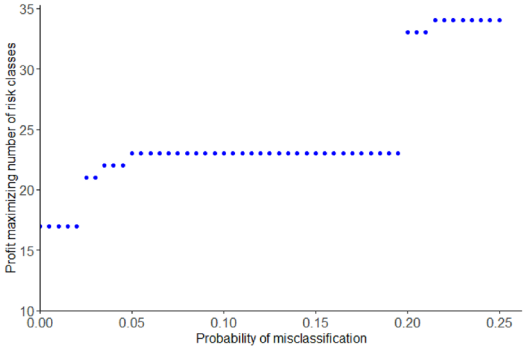
Results show that in the absence of classification costs, that is, the cost related to the setup and maintenance of an additional risk class, the insurer would opt for maximal granularity in classification, that is, offering 70 risk classes, regardless of the classification method used or level of initial misclassification. The same holds when using the default classification method and assuming a positive initial misclassification probability for any classification cost below a threshold $c \leq 30$. This means that the insurer has the incentive to treat each subpopulation as a separate risk class even if it does not have full information regarding the risk group to which policyholders belong. However, if we assume classification costs $c = 1000$ and an initial probability of misclassification $p = 15\%$, the maximal total profit is achieved by offering only 23 different risk classes. Panel (ii) in Fig. 3 shows the profit-maximizing number of risk classes depending on the assumed probability of misclassification. The cost of misclassification, that is, the difference between the theoretical maximal profit without misclassification and the maximal profit with misclassification, increases quickly with the number of risk classes, reaches its peak, and then decreases. This would suggest that, at first, the effect of increasing the total share of policyholders allocated to the incorrect risk class dominates—more (adjacent) risk classes lead to an overall higher share of total policyholders misclassified. However, as risk classes get more granular, the "missed" profit from each incorrectly classified risk becomes lower.

Could the insurer increase its profit by introducing a new classification methodology, using biological age, for instance, instead of age as risk factor? For calculating biological age, the insurer needs to use policyholders' private data and that, as discussed, will affect its demand function. If we assume that the insurer estimates no effect of using private data on client migration to/from competitors in either subpopulation, that is, $\beta^{70} = 1^{70}$, then the insurer would have an incentive to innovate using private data only if the compensation that its client base requires reflects in

[22] This choice of willingness to pay and cost data correspond to an equal target profitability in each risk group.



(i) Difference between profit in the innovative classification system versus in the default one.



(ii) Profit-maximizing number of risk classes for different levels of the probability of misclassification.

Fig. 3 Illustration of the profit-maximizing classification system in a heterogeneous population. This figure illustrates the selection of the profit-maximizing classification system in the case of a heterogeneous population. The underlying population is assumed to be composed of 70 subpopulations, with 100 (prospective) policyholders belonging to each subpopulation. The cost of maintaining an additional risk class is assumed to be 1000. The vector of maximal reservation prices per class decreases (linearly) from 3500 in the highest-risk subpopulation to 50 in the lowest-risk subpopulation, whereas the expected claim payment decreases (linearly) from 2800 in the highest-risk subpopulation to 40 in the lowest-risk subpopulation. Panel (i) displays the difference in (maximal) profit between using a classification system that yields a certain misclassification rate and using a classification system that eliminates the initial misclassification. The latter is enabled by using private data such that the demand for the new policy shifts with a certain expected α_1 and no change in the client base is expected, $\beta^{70} = 1^{70}$. The insurer would switch to the new classification methodology only for non-negative values of the differences in profit. Panel (ii) shows the number of risk classes that would maximize profits depending on different values of misclassification faced

$\alpha_1 = 0.986$.[23] In this case, it would be optimal for the insurer to offer only 17 risk classes instead of 23. If the base of potential policyholders is thought to be more concerned on average about the use of its (required) private data and, therefore, only

[23] Appendix 3 presents numerical results for alternative assumptions regarding the relationship between willingness to share private data and willingness to pay for insurance.

be expected to allow its use against a higher compensation, a profit-maximizing insurer would not innovate in risk classification using private data. Panel (i) in Fig. 3 shows the difference in (maximal) profits between the two classification methodologies for different combinations of the initial probability of misclassification (which can be corrected) and expected demand shift as measured by α_1.

Focusing on the effect that the use of policyholders' private data has on risk classification, our results do not fully validate the concern expressed in the literature that having more information on individual risk might lead to smaller risk pools (Eling and Lehmann 2018; Cevolini and Esposito 2020). In fact, for a given cost of setting up and maintaining an additional risk class, c, and a given (even small) fall in expected demand due to requiring private data, the optimal number of risk classes from the perspective of a profit-maximizing insurer in a fully accurate classification system is lower or equal to the optimal number of risk classes in a system with a positive probability of misclassification. A lower (or equal) number of risk classes is equivalent to the pooling of more (or as many) risk groups together in a risk class.

4 Outlook

While a large body of literature on big data, risk classification, and privacy in insurance markets has been emerging over recent years, we have identified a few avenues for future research that have not yet been (sufficiently) explored. To systematically identify those areas, we reviewed all outlook or future research sections in the literature referenced in Sect. 2. After excluding articles without future research sections and those published before 2011, a total of 41 articles that contain future research are analyzed. The review demonstrated that the topics of big data, risk classification, and privacy in insurance markets reach far beyond the field of economics. Therefore, we classify possible future research in the four categories economics, law, medicine, and ethics. Most of questions raised below are in the intersection of economics with other fields, emphasizing the need for cross-disciplinary research.

From the perspective of economic research, the current literature emphasizes the importance of privacy economics, but also points out that privacy protection is rapidly becoming a pressing public policy issue (Acquisti et al. 2016; Biener et al. 2020; Blakesley and Yallop 2020; Gemmo et al. 2020; Hoy and Durnin 2012; Loi et al. 2022; Steinberg 2022). Future research should improve our understanding of the economics of privacy and its interaction with insurance economics. While privacy preferences have been incorporated into several models, empirical research on the determinants of privacy preferences is still relatively scarce. Furthermore, as mentioned above, both the willingness to share private data and risk aversion are found to be influenced by similar consumer characteristics. This observation suggests that they are not independent of each other. The relationship between risk aversion and privacy concerns is particularly relevant to the study of insurance contracts, which require access to a wider range of private data. To our knowledge, this relationship has not been studied empirically in the insurance context.

In data and technology applications, future research could explore a broad range of topics. On the one hand, research should expand the scope of data sets, focusing specifically on driving behavior data to better understand customers' behavioral habits and improve the risk selection process (Baecke and Bocca 2017; Cather 2018; Biener et al. 2015; Brunnermeier et al. 2022). This can help insurance companies price insurance products more accurately and provide policies that better meet customer needs. On the other hand, the value and effect of sophisticated data mining techniques in risk selection should be further studied (Baecke and Bocca 2017; Holzapfel et al. 2023; Liu 2022, 2023). Additionally, the extensive use of big data highlights the importance of cyber insurance and requires for further research encompassing data testing and modeling, strategies to address information asymmetry in cyber risks, and the interplay between information asymmetry and network effects. The public good attributes of cybersecurity and the potential ramifications of government intervention also warrant further exploration (Biener et al. 2015; Eling and Lehmann 2018; Hassani et al. 2020; Xie et al. 2019). Among climate risk and global pandemics, Koijen and Yogo (2023) identify cyber risk as one of the *new risks*, for which the opportunities and challenges presented to the insurance industry offer interesting topics for future research. Big data and the related privacy considerations require an adequate data security protocol. With the ongoing process of digitalization and technological advancements, both the vehicles to protect sensitive data as well as those to breach this protection have become more developed. One challenge posed to insurance companies is that there are insufficient data available to accurately model the loss distribution for cyber risks (Koijen and Yogo 2023). While a substantial body of research currently focuses on utilizing insurance to mitigate cyber security risks (Biener et al. 2015; Bodin et al. 2018; Xie et al. 2019; Doss and Narasimhan 2021), there is a limited exploration of how the insurance sector responds to reputational risks and the ensuing cyber risk. A high level of uncertainty with respect to the loss distribution may increase premium loadings, increase deductibles, or decrease overall insurance supply, a result relevant for both cyber risk and reputational risk insurance. These circumstances may not only affect the demand for insurance, but the resulting lack of insurance coverage may alter household and firm behavior in the context of activities that exposes them to a high level of risk (Koijen and Yogo 2023).

For risks other than those that have emerged recently, new and more efficient ways to collect data allow for more accurate risk categorization. The aforementioned inverse selection dynamics, that is, the transfer of information advantages from the insured to the insurer (Villeneuve 2000, 2005; Brunnermeier et al. 2022), opens up a broad area for future research that revisits results from (standard and non-standard) asymmetric information models in which the information advantage has been on the side of the policyholder. In this paper, we propose a framework for an insurance company to navigate the decision process of whether and to what extent to implement new data-driven technologies. Our application analyzes an insurer's decision to choose the risk classification system that maximizes its expected profit. We consider the improvements in risk classification accuracy that can be achieved using big data, while taking into account the associated privacy implications. We do so by relating (potential) policyholders' willingness to provide additional private information to their willingness to pay for insurance. Our results suggest that improved risk classification accuracy, when achieved through the use of private data, does not necessarily lead to more granular risk classes. However,

reducing the cost of establishing and maintaining a separate risk class could lead to more granularity in risk classification.

Even when considering an informational advantage on the insurer's side, modeling policyholder behavior as well as insurance demand requires insurers to consider what level of information is available to households and to consider households' beliefs and preferences. For instance, a consumer's knowledge and beliefs about the loss distribution may diverge from the information the insurer matches to the respective consumer. In life and health insurance, consumers are likely to base their insurance demand on beliefs about their own health and longevity and they may or may not be able to estimate and consider their own biological age (Huang et al. 2017; Milevsky 2020a; Wu et al. 2021). The trust that consumers have in the insurance industry or individual firms (Courbage and Nicolas 2021; Gennaioli et al. 2022), their knowledge about their existing coverage, e.g., social insurance (Parente et al. 2005), as well as reliance on other safety nets (Kotlikoff and Spivak 1981; Brown et al. 2012) may also affect their willingness to pay for insurance. Empirical analyses of these mostly unobservable determinants of insurance demand, an exploration of correlations with observable characteristics, as well as theoretical models that incorporate such characteristics, could greatly help insurance companies in predicting insurance demand.

Other areas in which not a lot of research has been done are the potential changes in industrial organization which come along with the increasing use of big data. Additionally, the environmental cost of digital technologies and big data have been little explored in the literature (see Lucivero 2020; Samuel et al. 2022). It would be of interest to potentially quantify these costs in the insurance industry to get a clearer understanding of the trade-offs of utilizing big data.

There are numerous directions for future research in the intersection between insurance economics and legal. First, the application of big data will inevitably bring about the issue of insurability changes, which will trigger new legal issues (Eling and Kraft 2020). With the emergence of new possibilities of discrimination based on lifestyle tracking, new legal and regulatory challenges emerge (McFall 2019). Second, the legal and economic implications of self-insurance in the context of information asymmetry can be studied. This includes consideration of the impact on self-insurance of incomplete cure of disease, as well as the potential application of self-insurance in disease prevention. Some preventive measures may not be observable by insurance companies, which raises legal questions and ethical issues that require more in-depth research (Crainich 2017). This aspect is not only relevant in the health domain, but also more broadly in IT security. Further research is also needed to better understand the legal and economic challenges of risk scoring, paying particular attention to the legal issues multidimensional heterogeneity poses to the credit and insurance fields. As technology evolves, risk scoring models are likely to become more complex, requiring legal frameworks to accommodate new ways of using data (Einav et al. 2016; Eling and Lehmann 2018; Fuster et al. 2019; Hoy and Durnin 2012; Loi et al. 2022; Steinberg 2022). Also, the use of big data in fraud detection might require some interdisciplinary research on the legal barriers and economic implications of data usage. Overall, further economic and legal research is required to help inform the decision-makers in developing regulatory frameworks.

In the intersection with the sphere of medical research, future endeavors may further probe the application and impact of big data and genetic information in health insurance. This entails for example investigating the potential influence of genetic information on adverse selection and whether individuals with differing genetic test results adopt varying health insurance strategies (Crainich 2017; Filipova-Neumann and Hoy 2014; Hoy and Durnin 2012; Nill et al. 2019; Posey and Thistle 2021). Additionally, future research should delve more deeply into the utilization of various technologies in healthcare, encompassing the monitoring of medical applications of health data. These technologies offer the promise of enhanced disease prediction and prevention but also raise legal and ethical quandaries necessitating further research and regulation (Filipova-Neumann and Hoy 2014; Hoy and Durnin 2012; Nayak et al. 2019a, 2019b).

Lastly, future research must attend to the ethical dimensions of insurance, particularly given the continuous integration of new technologies. The insurance industry faces the challenge of balancing individual privacy and risk management needs (Biener et al. 2020; Meyers and van Hoyweghen 2020; Nill et al. 2019). Research should focus on examining the ethical repercussions of information asymmetry on data sharing and the ethical dilemmas arising from such asymmetry (McFall 2019). These areas of investigation will offer guidance for the future development of the insurance industry, ensuring that data and technology applications align with legal and ethical standards while meeting the expectations of customers and society (Aburto Barrera and Wagner 2023; Eling and Kraft 2020; Kiviat 2019; Tanninen et al. 2022; Wiegard and Breitner 2019).

Appendix 1

Review strategy and data collection

Our review methodology consists of three distinct phases, each with specific steps and inclusion criteria (Fig. 4). In the initial phase, we conducted a search using the Web of Science Core Collection database. In the first step of this phase (identification), we employed filters and keywords to refine our search within the database records. Our query encompassed all years up until December 2022, focusing solely on English documents and limiting the keyword search to the abstract. The selected keywords were carefully chosen to encapsulate the concept of big data in relation to insurance, with particular emphasis on economic, legal, medical, and ethical factors. These keywords are "insur*," "actuar*," "digital*," "big data," "mobile*," "classif*," "risk," "privacy," "legal," "medical*," and "ethical." The use of wildcard characters, denoted by the asterisk, allowed for variations of these terms to be captured. "Insur*" and "actuar*" ensured the inclusion of all publications pertaining to insurance or actuarial science, while "digital*," "big data," and "mobile*" filtered topics related to the digital landscape. Terms such as "classif*," "risk," "privacy," "legal," "medical*," and "ethical" were selected to specifically target risk classification, privacy concerns, and ethical considerations.[24] Through this systematic search process, we retrieved a total of 1089 publications.

[24] The exact query is: AB = (("insur*" OR "actuar*") AND ("digital*" OR "big data" OR "mobile*") AND ("classif*" OR "risk" OR "privacy" OR "legal" OR "medical*" OR "ethical")).

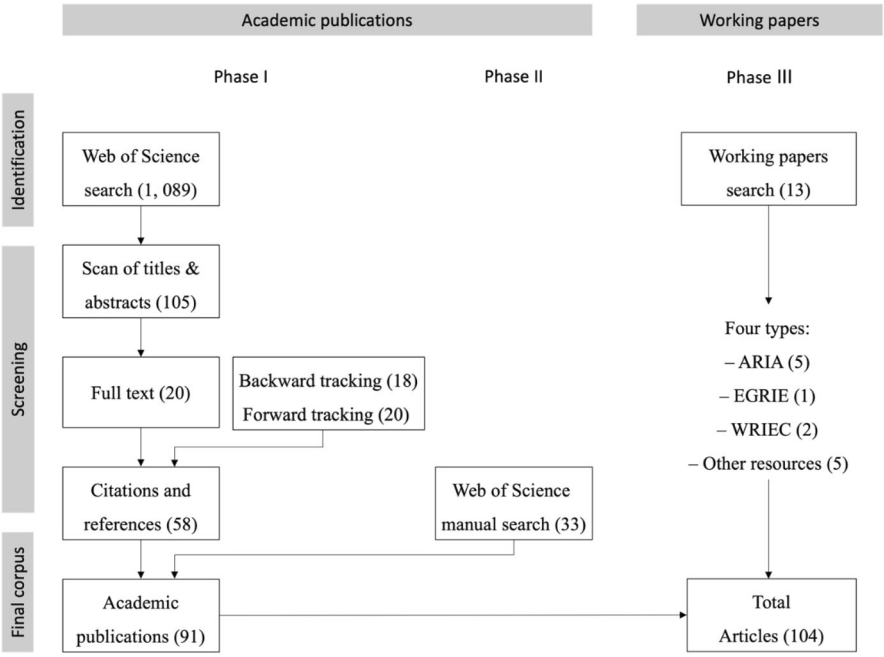


Fig. 4 Flow diagram for the identification and screening

In the second step (filtering), we deleted articles just mentioning "insurance can be a solution" instead of focusing on the topic of insurance. To do so we further added a query[25] based on the title of the article to limit the target articles to the scope of insurance. After this step, there are 105 articles left. We carefully examined the remaining records and excluded articles from fields of research that were not relevant to our study. We retained 28 of the 55 topics,[26] resulting in a set of 64 articles. The topic visualization for the original 105 articles is shown in Fig. 5. Most articles focus on medical treatment

[25] The exact query is: TI = ("insur" OR "actuar" OR "risk").

[26] In order to better identify articles that are closely related to the topics we discuss, we retain articles on the following 28 topics: Agricultural Policy, Artificial Intelligence and Machine Learning, Autonomic Regulation, Climate Change, Design and Manufacturing, Economic Theory, Economics, Environmental Sciences, Gender and Sexuality Studies, Genome Studies, Health Literacy and Telemedicine, Healthcare Policy, Homelessness and Human Trafficking, Information and Library Science, Law, Longevity, Management, Nursing, Nutrition and Dietetics, Ocean Dynamics, Safety and Maintenance, Security Systems, Software Engineering, Statistical Methods, Substance Abuse, Supply Chain and Logistics, Telecommunications, Transportation.

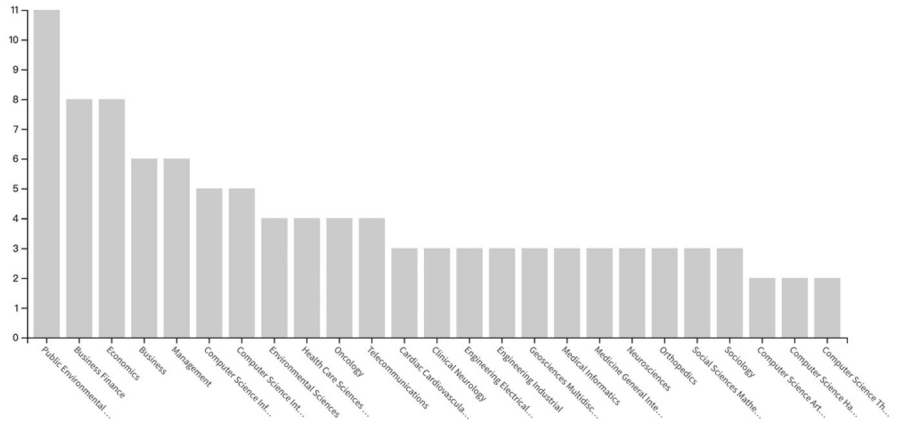


Fig. 5 Topic visualization

and health management; topics that were too focused on a particular type of disease had been excluded.[27]

In the third step (expansion), we screened the full texts of the remaining 64 articles, leading to the identification of 20 publications that met our inclusion criteria. Among the 64 articles involved, we considered the following two core criteria: first, whether the article is cited (i.e., citations in the web of science > 0) and whether the source of the literature is closely related to economics. We removed 12 articles that were not cited and 14 articles from journals on other topics (i.e., journals with themes such as criminology, feminism, and cerebrovascular diseases) from the 64 articles. We also deleted those articles whose source journals had an impact factor of 0, a total of 4 articles. Second, whether the article is closely related to the topic of insurance application big data. Due to the nature of the query we set, it may lead to the title and abstract of some articles only including the keyword "risk," without involving any words related to "insurance." This would take the subject of the article away from what we wanted to examine. There are also articles that incorrectly link to words that begin with "insur*," such as insurgent, insurgency, etc. For this reason, we excluded another 5 articles. Of the remaining articles, 5 focus on discussing how to use new computer methods to revolutionize the current risk classification method, and focus more on the description of the method, so we deleted it. Finally, there are 4 articles related to the description of the risk of the sick population, and the part related to insurance is to describe the patient as "uninsured." This is far from our topic, so we also deleted these articles. Therefore, we focus on 20

[27] These are: Allergy, Assisted Ventilation, Blood Clotting, Bone Diseases, Breast Cancer Scanning, Cardiac Arrhythmia, Cardiology-Circulation, Cardiology-General, Diabetes, Gastrointestinal and Esophageal Diseases, Hearing Loss, Hepatitis, Herbicides, Pesticides and Ground Poisoning, HIV, Lung Cancer, Microfluidic Devices and Superhydrophobicity, Neurodegenerative Diseases, Neuroscanning, Obstetrics and Gynecology, Oncology, Ophthalmology, Prostate Cancer, Reproductive Biology, Urology and Nephrology-General, Vascular, Cardiac and Thoracic Surgery, Virology-General, Wounds, and Ulcers.

articles centering on insurance and extending digitization and its economic and social consequences. These criteria focused on articles that positioned the insurance industry as a central player in the realm of big data and explored the economic and social consequences of its application within the industry. To ensure a comprehensive review, we also performed a forward and backward literature search for both citing and cited references related to the initially selected 20 records. This additional search resulted in the inclusion of a total of 18 citing and 20 cited articles, bringing the final count to 58 relevant records. Furthermore, to capture any insurance-related publications that might have been missed in the initial phase due to the specificity of our chosen keywords, we conducted a manual second phase of the search, identifying and including an additional 33 publications. Ultimately, our review encompassed a total of 91 academic publications in the field of big data in insurance. To also incorporate recent working papers, we manually review all papers from the annual meetings of the American Risk and Insurance Association (ARIA) from 2016 to 2022, the World Risk and Insurance Congress 2010, 2015, and 2020, and the European Group of Risk and Insurance Economists conferences from 2016 to 2022. Finally, we review citations in the identified working papers to explore additional relevant material. In addition, we search for the keywords same as above in the Social Science Research Network (SSRN) and via Google Scholar. We also identify numerous industry studies with these keywords by performing a regular Google search. Based upon this selection process, a database of 104 papers is set up and the main results are extracted.

Statistics on the corpus of academic literature

The 104 publications stem from 47 journals. Big Data and Society (8), The Geneva Papers on Risk and Insurance (7), Journal of Risk and Insurance (6), Journal of Health Economics (4), Journal of Business Ethics (3), and Risk Management and Insurance Review (3) are the journals with the highest number of articles that have published research on big data in insurance through 2022 (see Table 2). Of the 8 articles in the Big Data and Society, 5 studies discussed the relationship between insurance and big data itself, and 3 articles further discussed the economic and social consequences of applying big data in insurance.[28]

Only the top six journals are listed. The journals are ranked by number of records and listed in alphabetical order if equally ranked.

To identify and analyze the most relevant topics, we examined the frequency of keywords in the corpus from 2000 (oldest publication) to the end of 2022. In a first step, we report the most frequent topics based on the author's keywords field in the 104 publications (Fig. 6). To form the topics, we have clustered keywords with similar or related meanings. For example, the topic "insurance" includes the keywords

[28] The papers with the largest number of citations in the Web of Science (as of October 2, 2023) are Phelps et al. (2000; 530 citations), Bansal et al. (2010; 502 citations), Acquisti et al. (2016; 356 citations), Fuster et al. (2019; 218 citations), and Kehr et al. (2015; 194 citations). Many of the most cited papers fall in the field of privacy, emphasizing the general relevance of the topic not only in economics, but also beyond that.

Table 2 Journals with the highest number of publications in the final corpus

Journal	Number of records
Big Data and Society	8
Geneva Papers on Risk and Insurance	7
Journal of Risk and Insurance	6
Journal of Health Economics	4
Journal of Business Ethics	3
Risk Management and Insurance Review	3

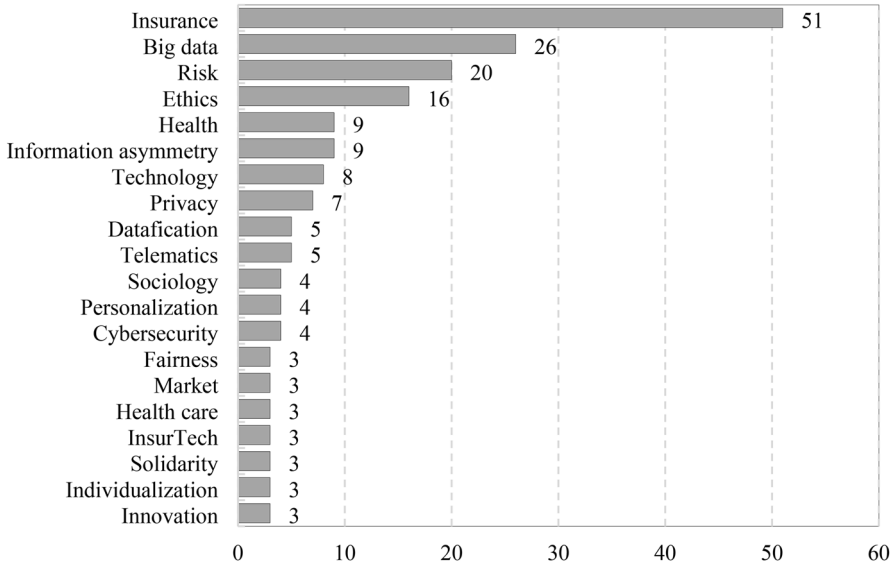


Fig. 6 Number of occurrences of the 20 most frequent topics in the keywords

insurance, insurer, insure, and insurers; "big data" includes data, big data, and big data analytics. The term "InsurTech" refers to Insurtech, InsurTechs, and Insurance Technology; "digitalization" includes the keywords of digitalization and digital words. Among the most frequent topics, we find that "insurance" ranks first with 51 occurrences. The topic "big data" ranks second with 26 occurrences. The topics "risk," and "ethics" rank third and fourth with 20 and 16 repetitions, respectively. The topics "health" and "information asymmetry" rank fifth with both nine repetitions. The frequency analysis of the 20 most frequent topics that we report in Fig. 2 provides insight into what has been of most interest to research over the past two decades. Besides "insurance," the keywords big data, risk, ethics, and health appear most frequently, which was to be expected given the search query for the selection of records. In the order of appearance, the application of big data is first, and then

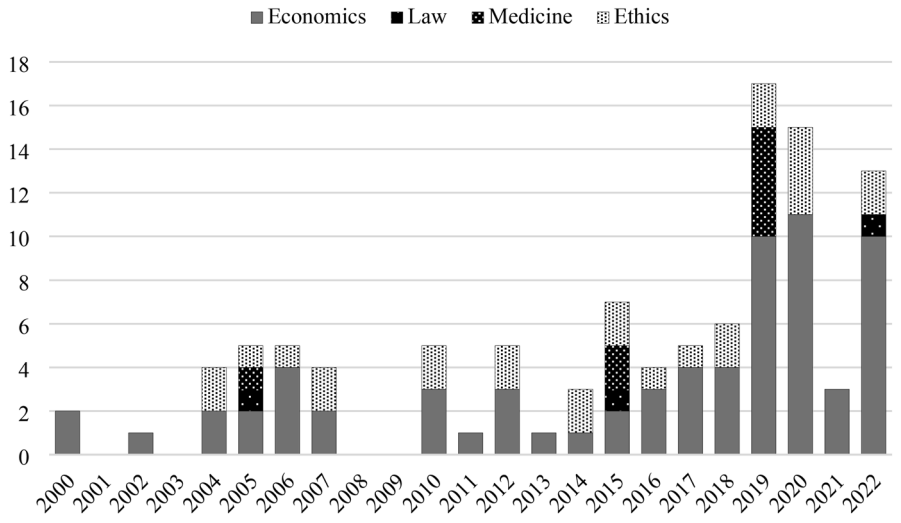


Fig. 7 Number of academic publications by field and year. *Note* a publication can be counted several times if related to several consequences

the technologies required to apply big data are considered. After that, these technologies are mainly applied to personalization and classification problems. We also observe that many keywords are related to the main characteristics of insurance such as risk, adverse selection, and InsurTech.

In a second step, we link the topics of each academic publication to two consequences of applying big data: economic benefits include updated insurance technologies and products, mitigation of information asymmetry problems, and expanded insurability. Social benefits consider topics related to legal, medical, ethical, and other issues affecting society. In Fig. 7, we show the number of recorded outcomes by consequence and year of applying big data in insurance. On the one hand, we observe that publications on economic factors appear in almost all years. Furthermore, the number of records appears to be increasing in (recent) years, from two record in 2000 to 10 records in 2022. On the other hand, regarding the discussion of the social consequences of the application of big data, most records occur between 2015 and 2022 (although there are also some records before 2010). We observe that in 2016 and previous studies, most of the studies discussed the possible digital application of a certain type of insurance product. After 2017, the research began to turn to the more general research on the application of big data in the insurance industry. This may be related to the concept of FinTech defined by the Financial Stability Board in 2016[29] and the National Association of Insurance Commissioners (NAIC) announced the creation of the Innovation and Technology Task Force[30] in 2017. Few publications have focused solely on the social consequences of applying big

[29] See https://www.fsb.org/work-of-the-fsb/financial-innovation-and-structural-change/fintech/.

[30] See https://naic.soutronglobal.net/Portal/Public/en-GB/RecordView/Index/24264.

data in insurance, perhaps because insurance has distinctly economic attributes in the first place. In the discussion of social benefits, research on medicine is relatively more popular. At the same time, concerns about ethics have always been there, and there has been relatively little research on law.

Overall, we observe increasing interest in studying the economic consequences of applying big data in insurance. Research on the social benefits brought by the application of big data in insurance has gradually increased in recent years and has involved more areas. For example, current research on the social benefits brought by the application of big data in insurance is gradually considering legal and regulatory factors and is attempting to regulate this emerging technology from a legal height to meet the needs of investors and consumers. At this stage, the statistics of the collected literature help us to get a first impression of the topics of most interest in academic research. Next, we will classify this literature under the topic of "big data for insurance applications" and provide an in-depth analysis of each category.

To examine the research areas covered by the retrieved literature corpus and to more systematically assess existing research and potential gaps, we conduct the analysis using the insurance value chain concept (Fig. 8). Two conceptual frameworks are used to present the results. The value chain distinguishes between the primary and supporting activities that a firm needs to deliver a product or service. Since Porter's value chain was formulated for general industry, we draw on the experience of Barrera and Wagner (2023) (see Fig. 4). We also rely on Principles on Artificial Intelligence (PoAI) from the National Association of Insurance Commissioners (NAIC). This is the NAIC's comprehensive regulatory guidance on insurers' use of consumer data and industry practices surrounding data technology. This guide includes five topics, namely Fair and Ethical, Accountable, Compliant, Transparent, and Secure, Safe, and Robust. The overarching principles of PoAI include strategies and operations for insurers to apply AI. Principles 2, 3, and 4 address the external stakeholders of the insurance industry, namely customers, suppliers, investors, governments, and regulators. In addition, the fifth principle deals with the process and risk management of insurers applying AI. We distinguish between supporting activities, core activities, and external stakeholders and reporting.

As illustrated in Fig. 4, we consider a framework based on nine main categories, including the value chain and relevant externalities. As a key representative

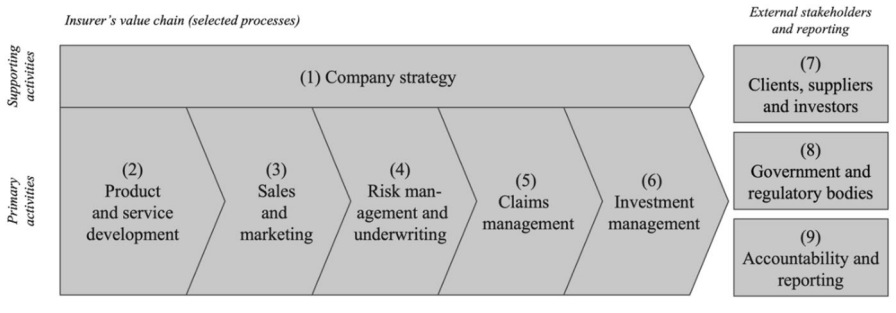


Fig. 8 Classification framework with nine categories along the value chain and stakeholders

for supporting activities in an insurance company we consider the company strategy (1). In the primary activities (operations), we consider product and service development (2), sales and marketing (3), risk management and underwriting (4), claims management (5), and investment management (6). Insurance companies are liable to several external stakeholders including clients, suppliers and investors (7), and the government and regulatory bodies (8) linked to their accountability and reporting (9). The proposed framework allows us to review which insurance activities are more researched (and concerned) with digitalization issues (see Table 3).

In Table 2, we report the number of publications of the final corpus that we have classified in each of the nine categories introduced in Fig. 8. Thereby a publication may refer to one or more categories. Additionally, in each category we consider the four factors to quantify the number of records relating to each result (Economics, Legal, Medicine, and Ethics). This split provides insights into which share of the extant literature covers these specific topics. The categories receiving the highest attention from academic research include product and service development (2), risk management and underwriting (4), and clients, suppliers, and investors (7). In each of these categories, we record over 30 publications, with most of them related to economic issues. The categories company strategy (1) and government and regulatory bodies (8) rank fourth and fifth in terms of the number of publications. All the other activities receive much less attention, in particular investment management (6) with merely four records. We observe that in most categories, academic research focuses primarily on economic results. In addition, we found that a significant proportion of research discusses the ethical consequences of applying digitalization at the risk management and underwriting stages, as well as at the customer service stage. The statistics reported in Table 2 highlight an important academic research gap, particularly in the categories of investment management, sales and marketing, and accountability and reporting where the number of publications is low. However, our statistics also show that the ethics issue has received the most attention over the past two decades, particularly regarding the category of clients. This may be

Table 3 Number of academic publications per category and corresponding results

Category	Number of records				
	Overall	Economics	Legal	Medicine	Ethics
(1) Company strategy	14	14	0	1	2
(2) Product and service development	38	37	2	6	10
(3) Sales and marketing	6	6	0	0	0
(4) Risk management and underwriting	50	48	4	6	21
(5) Claims management	8	7	2	2	2
(6) Investment management	4	4	0	0	0
(7) Clients, suppliers and investors	44	36	4	9	31
(8) Government and regulatory bodies	16	16	4	4	6
(9) Accountability and reporting	6	6	0	4	4

because the first application of insurance to big data is to apply customer data basically, which has sparked some academic discussions. Law results are less studied, although they get some attention in relation with government and regulatory bodies.

Appendix 2

Misclassification between two risk classes

We assume misclassification to run both directions, from the low-risk class to the high-risk class and vice versa. To better illustrate the net effect on total profit, we consider the simplest case of a population composed of two risk groups, high-risk and low risk, which the insurer classifies into two risk classes.

Prospective policyholders that belong to the high-risk group will be correctly classified into the high-risk class with a probability $1 - r < 0$ and incorrectly classified into the low-risk class with probability $r > 0$. Analogously, prospective policyholders that belong to the low-risk group will be correctly classified into the low-risk class with a probability $1 - r$ and incorrectly classified into the high-risk class with probability r. Figure 9 illustrates the demand curves and actuary cost curves in each risk group, showing the targeted profit-maximizing price-demand combinations as well as the additional demand created by misclassification.

Since we do not know specifically the willingness to pay/demand curve for the prospective policyholders that will be misclassified, we study the total profit breaking it down by risk group (as opposed to by risk class). That way the profit of the high-risk group can be expressed as

$$\pi_i = rn_{ij}\left(P_j^* - P_i^{\mathrm{A}}\right) + (1 - \mathrm{r})n_i^*\left(P_i^* - P_i^{\mathrm{A}}\right) \tag{1}$$

Equation (1) can be written as

$$\pi_i = r\left(N_i - \frac{N_i}{P_i^{\mathrm{R}}}P_j^*\right)\left(P_j^* - P_i^{\mathrm{A}}\right) + (1 - r)\left(N_i - \frac{N_i}{P_i^{R}}P_i^*\right)\left(P_i^* - P_i^{\mathrm{A}}\right), \tag{2}$$

where N_i stands for the total number of policyholders with a positive willingness to pay for insurance in risk group i, P_i^* denotes the targeted profit-maximizing price in risk group i, P_j^* denotes the targeted profit-maximizing price in risk group j, P_i^{A} denotes the actuarially fair premium in risk group i, and $P_i^{1\mathrm{R}}$ denotes the maximal reservation price of policyholders in risk group i.

The expected profit from insuring policyholders that belong to risk group j, that is, low risks, can be expressed analogously as

$$\pi_j = rn_{ji}\left(P_i^* - P_j^{\mathrm{A}}\right) + (1 - r)n_j^*\left(P_j^* - P_j^{\mathrm{A}}\right), \tag{3}$$

which can be written as

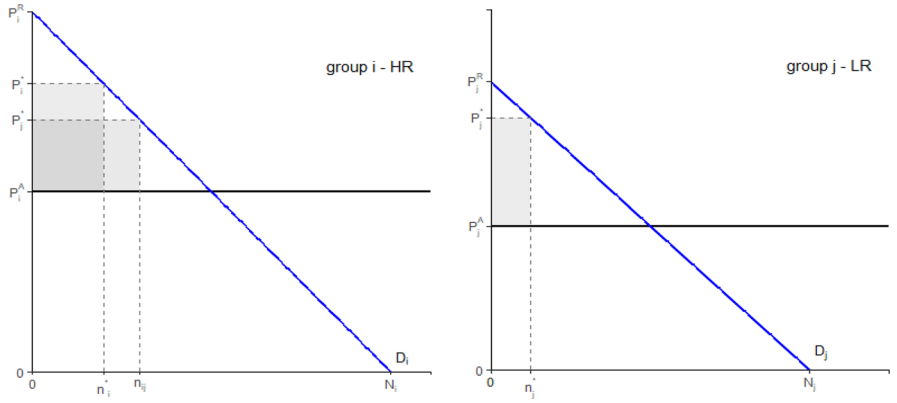


Fig. 9 The profit-maximizing targeted price and quantity in each risk group, modified for misclassification. This figure illustrates the expected profit-maximizing price–quantity for the insurer, given the demand that it faces from each group and the change in the expected maximum profit given the misclassification of policyholders from group i as group j. P_i^{R} and P_j^{R} denote the highest willingness to pay for insurance among policyholders in groups i and j, respectively. P_i^{A} and P_j^{A} stand for the expected indemnity payment per policyholder in groups i and j, respectively. P_i^* and P_j^* denote the profit-maximizing targeted price that the insurer sets in groups i and j, respectively. N_i and N_j represent the total number of (prospective) policyholders with a positive willingness to pay for insurance in groups i and j, respectively. n_i^* and n_j^* denote the respective number of policyholders in groups i and j that buy insurance at the price P_i^* and P_j^*, respectively. n_{ij} represents the additional demand created by misclassifying (prospective) policyholders of group i into group j and offering them insurance at the price P_j^*. Note that (prospective) policyholders of group j that are misclassified into group i are offered the price P_i^*, which does not translate into a positive demand in the demand curve of group j. Therefore, this flow of misclassified policyholders is not depicted in the graph

$$\pi_j = r\left(N_j - \frac{N_j}{P_j^{\mathrm{R}}}P_i^*\right)\left(P_i^* - P_j^{\mathrm{A}}\right) + (1-r)\left(N_j - \frac{N_j}{P_j^{\mathrm{R}}}P_j^*\right)\left(P_j^* - P_j^{\mathrm{A}}\right), \qquad (4)$$

where r stand for the initial probability of misclassification from risk group j to i (to the advantage of the insurer); P_j^{A} is the actuarially fair premium in risk group j, and N_j stands for the total number of policyholders that belong to risk group j. Note that the term $N_j - \frac{N_j}{P_j^{\mathrm{R}}}P_i^*$ in Eq. (4) represents the additional demand, n_{ji}, created by offering to misclassified policyholders from group j coverage for a price P_i^*. Therefore, it must be non-negative. In the case of the misclassification from low risk to high-risk, it is more likely for the additional demand created by misclassification to be zero or very small due to possibly lower willingness to pay, as illustrated in the case shown in Fig. 9.

Either way, whether the individual misclassification is to the advantage of the insurer or not, we can see that a percentage r of prospective policyholders in the risk group(s) will/might have a resulting price–demand combination that deviates from the profit-maximizing one. Therefore, the total profit will be lower than in the absence of misclassification.

Numerical implementation to term life insurance

Table 4 Variable names and definitions, and indexes overview

Variables	Variable name	Description
s	Risk group index	Index referring to a given risk group
P_s^{1R}	Highest (initial) reservation price in the risk group s	The highest reservation price from (prospective) policyholders' perspective in the risk group s, for the initial policy (not using private data)
P_s^{A}	Expected claim payment per policy in the risk group s	The expected claim payment per policy in the risk group s; assumed constant and equal to the average cost per unit of a policy in the group. Does not change regardless of the classification system used
N_s^1	(Initial) number of policyholders in the risk group s	(Initial) number of (prospective) policyholders with a positive willingness to pay in the risk group s
α_1	Coefficient of demand shift in the first scenario	Coefficient of fall in the highest reservation price due to compensation required for the use of private data, when willingness to pay for insurance and willingness to share private data are negatively correlated; assumed lower than 1
α_2	Coefficient of demand shift in the second scenario	Coefficient of fall in the highest reservation price due to compensation required for the use of private data, when willingness to pay for insurance and willingness to share private data are positively correlated; assumed lower than 1
α_3	Coefficient of demand shift in the third scenario	Coefficient of the parallel shift in demand curve due to compensation required for the use of private data, when willingness to pay for insurance and willingness to share private data are not correlated; assumed lower than 1
β_s	Coefficient of demand shift from client migration in the risk group s	Coefficient of demand shift from client migration in the risk group s
c	Classification cost per additional class	Classification cost per additional class
ur	Probability of misclassification	Probability of misclassifying perspective policyholders between adjacent risk classes
m	Classification System	The classification system, as defined by the number of risk classes and the combination of the probability of misclassification (r or 0) and compensation for private data usage (0 or $x > 0$). x is determined by the demand shift coefficients
I_m	Number of risk classes	The number of risk classes in a classification system
d	Index of the default classification methodology	The superscript d indicates variables under the default classification methodology
i	Index of the innovative classification methodology	The superscript n indicates variables under the innovative classification methodology

Table 5 Profit of each possible classification (with and without using private data) in a population composed of five (risk type) subpopulations

Number of classes	(Hypothetical) profit assuming classification is fully accurate	Profit accounting for misclassification $p_{ij}^1 = 0.2$	Profit of fully accurate classification, with $\alpha_1 = 0.946$.
1	2365.30	2365.30	2177.59
2	3213.67	2734.73	2954.01
3	3176.42	2845.61	2923.83
4	3164.54	2953.89	2905.32
5	3083.12	2923.17	2824.36

This table presents the profit for an insurer that faces an underlying policyholders' population composed of five subpopulations (in terms of risk type) under several risk classification systems. The subpopulations are considered to have the demand and cost characteristics of the five age groups of smokers from Braun et al. (2016). The linear approximation of the demand curve of each group is considered. The risk classification systems are characterized by the number of classes, shown in column 1, the misclassification rate, and the expected shift in the demand curve when using private data. The misclassification rate is assumed to be 0.2 in the default case (demand curve as the linear approximation from Braun et al. 2016) and zero when using private data in the latter case a demand shift characterized by $\alpha_1 = 0.946$ is expected. The cost of maintaining one additional risk class is assumed $c = 100$. Column 2 presents a hypothetical profit calculated for each number of risk classes, in the absence of misclassification and with no shift in demand. Column 3 presents the profit calculated for each number of risk classes, using the default classification methodology that yields a misclassification rate $p_{ij}^1 = 0$. Column 3 presents the profit calculated for each number of risk classes, using the new classification methodology that eliminates misclassification but is associated with a shift in demand characterized by $\alpha_1 = 0.946$. The profit for the number of risk classes that would maximize profits in each classification methodology is highlighted (Tables 4, 5, 6).

Table 6 Profit-maximizing number of risk classes under different classification methodologies, when the willingness to pay for insurance and the willingness to share private data are negatively related

[1]	[2]	[3]	[4]	[5]
2000	0.00%	14	1.0000	14
	2.50%	14	0.9965	14
	5.00%	14	0.9930	14
	7.50%	14	0.9894	14
	10.00%	14	0.9857	14
	12.50%	17	0.9824	14
	15.00%	17	0.9791	14
	17.50%	17	0.9757	14
	20.00%	23	0.9728	14
	22.50%	23	0.9704	14
	25.00%	23	0.9679	14
1000	0.00%	17	1.0000	17
	2.50%	21	0.9969	17
	5.00%	23	0.9945	17
	7.50%	23	0.9922	17
	10.00%	23	0.9901	17
	12.50%	23	0.9878	17
	15.00%	23	0.9856	17
	17.50%	23	0.9833	17
	20.00%	33	0.9810	17
	22.50%	34	0.9795	17
	25.00%	34	0.9781	17
500	0.00%	22	1.0000	22
	2.50%	23	0.9978	22
	5.00%	31	0.9960	22
	7.50%	33	0.9944	22
	10.00%	34	0.9930	22
	12.50%	34	0.9917	22
	15.00%	34	0.9904	22
	17.50%	35	0.9890	22
	20.00%	35	0.9877	22
	22.50%	35	0.9864	22
	25.00%	35	0.9851	22
200	0.00%	32	1.0000	32
	2.50%	35	0.9986	32
	5.00%	35	0.9974	32
	7.50%	35	0.9961	32
	10.00%	52	0.9950	32
	12.50%	55	0.9940	32
	15.00%	58	0.9932	32
	17.50%	61	0.9924	32
	20.00%	64	0.9917	32
	22.50%	67	0.9910	32
	25.00%	69	0.9904	32
50	0.00%	55	1.0000	55
	2.50%	67	0.9993	55

Table 6 (continued)

[1]	[2]	[3]	[4]	[5]
	5.00%	69	0.9987	55
	7.50%	69	0.9981	55
	10.00%	69	0.9975	55
	12.50%	69	0.9969	55
	15.00%	70	0.9963	55
	17.50%	70	0.9958	55
	20.00%	70	0.9952	55
	22.50%	70	0.9946	55
	25.00%	70	0.9940	55
30	0.00%	69	1.0000	69
	2.50%	70	0.9994	69
	5.00%	70	0.9988	69
	7.50%	70	0.9983	69
	10.00%	70	0.9977	69
	12.50%	70	0.9971	69
	15.00%	70	0.9965	69
	17.50%	70	0.9959	69
	20.00%	70	0.9953	69
	22.50%	70	0.9947	69
	25.00%	70	0.9942	69
0	0.00%	70	1.0000	70
	2.50%	70	0.9994	70
	5.00%	70	0.9988	70
	7.50%	70	0.9983	70
	10.00%	70	0.9977	70
	12.50%	70	0.9971	70
	15.00%	70	0.9965	70
	17.50%	70	0.9959	70
	20.00%	70	0.9953	70
	22.50%	70	0.9947	70
	25.00%	70	0.9942	70

This table presents the profit-maximizing number of risk classes under a classification methodology that classifies risks with a given probability of misclassification versus the profit-maximizing number of risk classes under a classification methodology that classifies risks fully accurately. The classification methodology that classifies risks fully accurately is enabled using private data, which in turn, is associated with a shift of the demand curve according to the scenario presented in Fig. 2 in the paper. Column 1 presents the classifications cost increment, that is, the costs related to setting up and maintaining an additional risk class; Column 2 presents values of the probability of misclassification under the classification methodology that classifies risks with a given probability of misclassification; Column 3 presents the profit-maximizing number of risk classes when using the classification methodology that classifies with a given probability of misclassification; Column 4 presents the maximal demand shift coefficient α_1 for which the insurer would be indifferent between using the old classification method or a new one that fully eliminates misclassification; Column 5 presents the profit-maximizing number of risk classes when using the new classification method that is fully accurate and when demand is shifted by α_1.

Appendix 3

Alternative scenarios for the relationship between willingness to pay for insurance and privacy concerns

This Appendix discusses two alternative assumptions with respect to the relationship between willingness to pay for insurance and privacy concerns and how they would impact our analyses in Sect. 3.1. Considering different scenarios with respect to the said relationship is an approach supported by the theory of domain-specific risk aversion (Blais und Weber 2006). In that context, privacy concerns might be categorized into different domains depending on the line of business and the type of data required. Below, we denote these alternative scenarios as scenario (b) and (c) (implicitly referring to the scenario discussed in Sect. 3.1 as scenario (a)).

In scenario (b), we assume the willingness to pay for insurance to be positively related to the willingness to share data. This would mean that the policyholders with the highest willingness to pay for insurance would be willing to share data with the insurer for a minimal reduction in price, whereas those with a lower willingness to pay for insurance would request a higher deduction in price compared to the price they would be willing to pay for the initial policy. In addition, to pin down the shape of the demand shift, we assume that policyholders with higher willingness to pay for insurance coverage do not require any compensation for sharing their private data; therefore, their willingness to pay for either product is the same. The left graph in Fig. 10 depicts the demand shift in a given risk group.

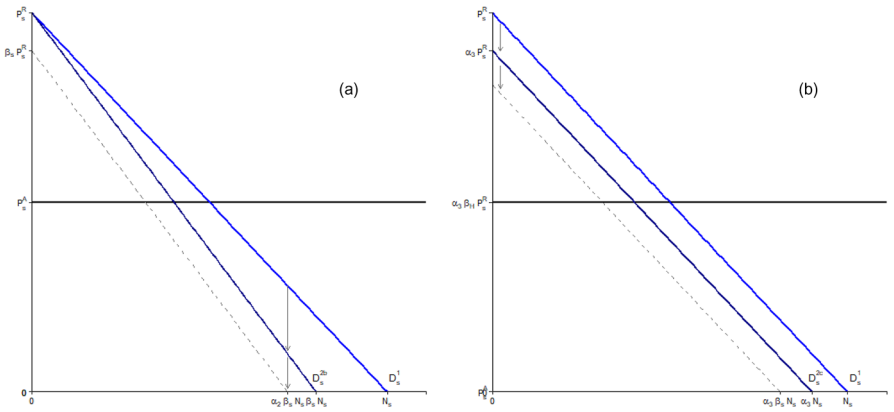


Fig. 10 Demand curve shift based on the relationship between willingness to share private data and willingness to pay for insurance. This figure illustrates the shifts in demand curve for the new insurance policy in each risk group, under the alternative assumptions regarding the relation between the willingness to pay for insurance and the willingness to share private data. **a** Depicts the case where the willingness to pay for insurance is positively related to the willingness to share data; and **b** depicts the case where the willingness to pay for insurance is independent from the willingness to share data. The magnitude of the demand shift is quantified through the coefficients α_k, for $k = 2, 3$, that reflects the shift due to the required compensation for sharing data

In scenario (c), we assume the willingness to pay for insurance and willingness to share private data to be independent. In this case, the demand curve would shift downward parallelly and the request to share private data would have the same effect as an increase in cost per policy. The right graph in Fig. 10 depicts the demand shift in a given risk group. The parallel demand shift due to the compensation required for sharing data is described by the coefficient, α_3.

In what follows, we express the expected profit under the new classification methodology based on the alternative assumptions regarding the demand curve shift. To this end, we will denote the profits in each risk group from (10) with the superscript b and c in the alternative scenarios (b) and (c), respectively.

Note that the definition of the demand curve shift coefficient, α_k for $k = 1, 2, 3$, is slightly different in each scenario to simplify the expression of profit for the new product. In either case, α_k can be used to express the new set of prices that maximizes the insurer's expected profit. Using this notation simplifies the analysis of the magnitude of the demand shift for which the new maximal profit in each group, coupled with full classification accuracy, would lead to higher total profits.

In each of these scenarios, to estimate the maximal value of demand shift, characterized by α_k, $k = 1, 2, 3$, for which the new policy is at least as profitable as the initial one, we set $\pi^1 = \pi^2$. This way, the insurer can calculate the maximal demand shift for which innovation in risk classification methods using private data would increase the expected profit. Comparing the resulting demand with the one estimated in the market would lead to a decision on whether to offer the new policy that requires access and permission to use private data.

A numerical example of a population with 70 homogenous subpopulations under an alternative scenario

This section provides a numerical example analog to the one described in Sect. 3.3.2, with the difference that the willingness to pay for insurance and the willingness to share private data are assumed to be positively related according to scenario (b). Table 7 presents the combinations of the demand shift α_2 and the profit-maximizing number of risk classes that would make the insurer indifferent between implementing a risk classification methodology that fully eliminates misclassification and keeping the 'old' risk classification system. Results are presented for a range of values of the initial misclassification, for a certain cost of maintaining an additional risk class. For instance, results show that under this scenario the maximal demand shift for which the insurer would innovate for correcting a 15% probability of misclassification given a cost of maintaining an additional risk class of $c = 1000$ and no expected shift due to client migration, that is, $\beta^{70} = 1^{70}$, corresponds to $\alpha_2 = 0.8631$. Furthermore, even under this scenario, improved accuracy does not give the insurer an incentive to form more granular risk classes.

Table 7 Profit-maximizing number of risk classes under different classification methodologies, when the willingness to pay for insurance and the willingness to share private data are negatively related

[1]	[2]	[3]	[4]	[5]
2000	0.00%	14	1.0000	14
	2.50%	14	0.9652	14
	5.00%	14	0.9302	13
	7.50%	14	0.8947	13
	10.00%	14	0.8593	13
	12.50%	17	0.8278	13
	15.00%	17	0.7963	13
	17.50%	17	0.7648	13
	20.00%	23	0.7392	13
	22.50%	23	0.7159	11
	25.00%	23	0.6924	10
1000	0.00%	17	1.0000	17
	2.50%	21	0.9704	17
	5.00%	23	0.9467	17
	7.50%	23	0.9258	17
	10.00%	23	0.9049	17
	12.50%	23	0.8840	17
	15.00%	23	0.8631	17
	17.50%	23	0.8422	17
	20.00%	33	0.8218	17
	22.50%	34	0.8085	17
	25.00%	34	0.7956	16
500	0.00%	22	1.0000	22
	2.50%	23	0.9796	22
	5.00%	31	0.9623	22
	7.50%	33	0.9478	22
	10.00%	34	0.9351	22
	12.50%	34	0.9228	22
	15.00%	34	0.9104	22
	17.50%	35	0.8984	22
	20.00%	35	0.8865	22
	22.50%	35	0.8746	22
	25.00%	35	0.8627	22
200	0.00%	32	1.0000	32
	2.50%	35	0.9877	32
	5.00%	35	0.9761	32
	7.50%	35	0.9646	31
	10.00%	52	0.9539	31
	12.50%	55	0.9454	31
	15.00%	58	0.9378	31
	17.50%	61	0.9307	31
	20.00%	64	0.9243	31
	22.50%	67	0.9183	31
	25.00%	69	0.9129	31
50	0.00%	55	1.0000	55
	2.50%	67	0.9934	55

Table 7 (continued)

[1]	[2]	[3]	[4]	[5]
	5.00%	69	0.9881	55
	7.50%	69	0.9829	55
	10.00%	69	0.9776	55
	12.50%	69	0.9724	55
	15.00%	70	0.9672	54
	17.50%	70	0.9619	54
	20.00%	70	0.9567	54
	22.50%	70	0.9515	54
	25.00%	70	0.9463	54
30	0.00%	69	1.0000	69
	2.50%	70	0.9948	68
	5.00%	70	0.9896	68
	7.50%	70	0.9844	68
	10.00%	70	0.9792	68
	12.50%	70	0.9740	68
	15.00%	70	0.9688	67
	17.50%	70	0.9636	67
	20.00%	70	0.9584	67
	22.50%	70	0.9532	67
	25.00%	70	0.9480	67
0	0.00%	70	1.0000	70
	2.50%	70	0.9948	70
	5.00%	70	0.9896	70
	7.50%	70	0.9844	70
	10.00%	70	0.9792	70
	12.50%	70	0.9740	70
	15.00%	70	0.9688	70
	17.50%	70	0.9636	70
	20.00%	70	0.9584	70
	22.50%	70	0.9532	70
	25.00%	70	0.9480	70

This table presents the profit-maximizing number of risk classes under a classification methodology that classifies risks with a given probability of misclassification versus the profit-maximizing number of risk classes under a classification methodology that classifies risks fully accurately. The classification methodology that classifies risks fully accurately is enabled using private data, which in turn is associated with a shift of the demand curve following the scenario presented in panel (i) of Fig. 10. Column 1 presents the classifications cost increment, that is, the costs related to setting up and maintaining an additional risk class; Column 2 presents values of the probability of misclassification under the classification methodology that classifies risks with a given probability of misclassification; Column 3 presents the profit-maximizing number of risk classes when using the classification methodology that classifies with a given probability of misclassification; Column 4 presents the maximal demand shift coefficient α_2, for which the insurer would be indifferent between using the old classification method or a new one that fully eliminates misclassification; Column 5 presents the profit-maximizing number of risk classes when using the new classification method that is fully accurate and when demand is shifted by α_2

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1057/s10713-024-00098-5.

Acknowledgements We thank the participants of the 2022 and 2023 EGRIE Seminar, 2022 ARIA Annual Meeting, the 2022 Journees Internationales du Risque (JIR) and, PiF Seminar at the University of St. Gallen for questions and comments. We thank the discussants Julia Holzapfel, Xin Che, and Thomas Maurice for their valuable feedback. Special thanks to Tongpu Zhao for research assistance in preparing the literature review and outlook.

Funding Open access funding provided by University of St.Gallen.

References

Aburto Barrera, L.I., and J. Wagner. 2023. A systematic literature review on sustainability issues along the value chain in insurance companies and pension funds. *European Actuarial Journal* 13: 1–49.

Acquisti, A., C. Taylor, and L. Wagman. 2016. The economics of privacy. *Journal of Economic Literature* 54 (2): 442–492.

Albrecher, H., A. Bommier, D. Filipović, P. Koch-Medina, S. Loisel, and H. Schmeiser. 2019. Insurance: Models, digitalization, and data science. *European Actuarial Journal* 9: 349–360.

Altman, D., D.M. Cutler, and R.J. Zeckhauser. 1998. Adverse selection and adverse retention. *The American Economic Review* 88 (2): 122–126.

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

Balasubramanian, R., A. Libarikian, and D. McElhaney. 2018. *Insurance 2030—The impact of AI on the future of insurance*. New York: McKinsey and Company.

Bansal, G., F. Zahedi, and D. Gefen. 2010. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision Support Systems* 49 (2): 138–150.

Barigozzi, F., and D. Henriet. 2011. Genetic information: Comparing alternative regulatory approaches when prevention matters. *Journal of Public Economic Theory* 13 (1): 23–46.

Barry, L. 2020. Insurance, big data and changing conceptions of fairness. *European Journal of Sociology/archives Européennes De Sociologie* 61 (2): 159–184.

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

Bednarz, Z., and K. Manwaring. 2022. Hidden depths: The effects of extrinsic data collection on consumer insurance contracts. *Computer Law and Security Review* 45: 105667.

Bélisle-Pipon, J.-C., E. Vayena, R.C. Green, and I.G. Cohen. 2019. Genetic testing, insurance discrimination and medical research: What the United States can learn from peer countries. *Nature Medicine* 25 (8): 1198–1204.

Benndorf, V., and H.-T. Normann. 2018. The willingness to sell personal data. *The Scandinavian Journal of Economics* 120 (4): 1260–1278.

Biener, C., M. Eling, and J.H. Wirfs. 2015. Insurability of cyber risk: An empirical analysis. *The Geneva Papers on Risk and Insurance: Issues and Practice* 40: 131–158.

Biener, C., M. Eling, and M. Lehmann. 2020. Balancing the desire for privacy against the desire to hedge risk. *Journal of Economic Behavior and Organization* 180: 608–620.

Blais, Ann-Renée., and Elke U. Weber. 2006. A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making* 1 (1): 33–47.

Blakesley, I.R., and A.C. Yallop. 2019. What do you know about me? Digital privacy and online data sharing in the UK insurance sector. *Journal of Information, Communication and Ethics in Society* 18 (2): 281–303.

Blakesley, I.R., and A.C. Yallop. 2020. What do you know about me? Digital privacy and online data sharing in the UK insurance sector. *Journal of Information, Communication and Ethics in Society* 18 (2): 281–303.

Blasimme, A., E. Vayena, and I. van Hoyweghen. 2019. Big Data, precision medicine and private insurance: A delicate balancing act. *Big Data and Society* 6 (1): 2053951719830111.

Bodin, L.D., L.A. Gordon, M.P. Loeb, and A. Wang. 2018. Cybersecurity insurance and risk-sharing. *Journal of Accounting and Public Policy* 37 (6): 527–544.

Bohnert, A., A. Fritzsche, and S. Gregor. 2019. Digital agendas in the insurance industry: The importance of comprehensive approaches. *The Geneva Papers on Risk and Insurance: Issues and Practice* 44: 1–19.

Bologa, A.-R., R. Bologa, and A. Florea. 2013. Big Data and specific analysis methods for insurance fraud detection. *Database Systems Journal* 4 (4): 30–39.

Bond, E.W., and K.J. Crocker. 1991. Smoking, skydiving, and knitting: The endogenous categorization of risks in insurance markets with asymmetric information. *Journal of Political Economy* 99 (1): 177–200.

Braun, A., H. Schmeiser, and F. Schreiber. 2016. On consumer preferences and the willingness to pay for term life insurance. *European Journal of Operational Research* 253 (3): 761–776.

Braun, A., N. Haeusle, and P. Thistle. 2023. Risk classification with on-demand insurance. *Journal of Risk and Insurance* 90 (4): 975–990.

Breidbach, C.F., and P. Maglio. 2020. Accountable algorithms? The ethical implications of data-driven business models. *Journal of Service Management* 31 (2): 163–185.

Brown, J.R., G.S. Goda, and K. McGarry. 2012. Long-term care insurance demand limited by beliefs about needs, concerns about insurers, and care available from family. *Health Affairs* 31 (6): 1294–1302.

Browne, M.J., and S. Kamiya. 2012. A theory of the demand for underwriting. *Journal of Risk and Insurance* 79 (2): 335–349.

Brunnermeier, M.K., R. Lamba, and C. Segura-Rodriguez. 2022. Inverse selection. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3584331.

Castillo, M.J., S. Boucher, and M. Carter. 2016. Index insurance: Using public data to benefit small-scale agriculture. *International Food and Agribusiness Management Review* 19 (1030-2016–83144): 93–114.

Cather, D.A. 2018. Cream skimming: Innovations in insurance risk classification and adverse selection. *Risk Management and Insurance Review* 21 (2): 335–366.

Cesarini, L., R. Figueiredo, B. Monteleone, and M.L.V. Martina. 2021. The potential of machine learning for weather index insurance. *Natural Hazards and Earth System Sciences* 21 (8): 2379–2405.

Cevolini, A., and E. Esposito. 2020. From pool to profile: Social consequences of algorithmic prediction in insurance. *Big Data and Society* 7 (2): 205395172093922.

Charpentier, A., L. Barry, and M.R. James. 2022. Insurance against natural catastrophes: Balancing actuarial fairness and social solidarity. *The Geneva Papers on Risk and Insurance: Issues and Practice* 47 (1): 50–78.

Che, X., A. Liebenberg, and J. Xu. 2022. Usage-based insurance—Impact on insurers and potential implications for InsurTech. *North American Actuarial Journal* 26 (3): 428–455.

Ciborra, C. 2006. Imbrication of representations: Risk and digital technologies. *Journal of Management Studies* 43 (6): 1339–1356.

Courbage, C., and C. Nicolas. 2021. Trust in insurance: The importance of experiences. *Journal of Risk and Insurance* 88 (2): 263–291.

Crainich, D. 2017. Self-insurance with genetic testing tools. *Journal Risk and Insurance* 84 (1): 73–94.

Crocker, K.J., and A. Snow. 2000. The theory of risk classification. In *Handbook of insurance*. Huebner international series on risk, insurance, and economic security, ed. J.D. Cummins, and G. Dionne, 245–276. Dordrecht: Springer.

Crocker, K.J., and N. Zhu. 2021. The efficiency of voluntary risk classification in insurance markets. *Journal of Risk and Insurance* 88 (2): 325–350.

Dionne, G., and C. Rothschild. 2014. Economic effects of risk classification bans. *The Geneva Risk and Insurance Review* 39 (2): 184–221.
Doherty, N.A., and L.L. Posey. 1998. On the value of a checkup: Adverse selection, moral hazard and the value of information. *Journal of Risk and Insurance* 65 (2): 189.
Doss, S., and R. Narasimhan. 2021. Qualitative assessment of cyber risk exposures in India. *Asia-Pacific Journal of Risk and Insurance* 15 (2): 85–105.
Einav, L., and J. Levin. 2014. Economics in the age of big data. *Science* 346 (6210): 1243089.
Einav, L., A. Finkelstein, and J. Levin. 2010. Beyond testing: Empirical models of insurance markets. *Annual Review of Economics* 2: 311–336.
Einav, L., A. Finkelstein, R. Kluender, and P. Schrimpf. 2016. Beyond statistics: The economic content of risk scores. *American Economic Journal. Applied Economics* 8 (2): 195–224.
Eling, M., and M. Kraft. 2020. The impact of telematics on the insurability of risks. *The Journal of Risk Finance* 21 (2): 77–109.
Eling, M., and M. Lehmann. 2018. The impact of digitalization on the insurance value chain and the insurability of risks. *The Geneva Papers on Risk and Insurance: Issues and Practice* 43 (3): 359–396.
Eling, M., R. Jia, J. Lin, and C. Rothschild. 2022. Technology heterogeneity and market structure. *Journal of Risk and Insurance* 89 (2): 427–448.
Fang, H., X. Qin, W. Wu, and T. Yu. 2020. Mutual risk sharing and Fintech: The case of Xiang Hu Bao. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3781998.
Farboodi, M., D. Singal, L. Veldkamp, and V. Venkateswaran. 2022. *Valuing financial data*. Cambridge: NBER.
Farrell, J. 2012. Can privacy be just another good. *Journal on Telecommunication and High Technology Letters* 10: 251.
Faure, M., and S. Li. 2020. Risk shifting in the context of 3D printing: An insurability perspective. *The Geneva Papers on Risk and Insurance: Issues and Practice* 45: 482–507.
Filipova, L. 2006. *Endogenous information and privacy in automobile insurance markets*. BGPE Discussion Paper.
Filipova, L. 2007. *Monitoring and privacy in automobile insurance markets with moral hazard*. BGPE Discussion Paper No. 26.
Filipova-Neumann, L., and M. Hoy. 2014. Managing genetic tests, surveillance, and preventive medicine under a public health insurance system. *Journal of Health Economics* 34: 31–41.
Filipova-Neumann, L., and P. Welzel. 2010. Reducing asymmetric information in insurance markets: Cars with black boxes. *Telematics and Informatics* 27 (4): 394–403.
Francois, P., and T. Voldoire. 2022. The revolution that did not happen: Telematics and car insurance in the 2010s. *Big Data and Society* 9 (2): 205395172211420.
Fritzsch, S., P. Scharner, and G. Weiß. 2021. Estimating the relation between digitalization and the market value of insurers. *Journal of Risk and Insurance* 88 (3): 529–567.
Fuster, A., M. Plosser, P. Schnabl, and J. Vickery. 2019. The role of technology in mortgage lending. *The Review of Financial Studies* 32 (5): 1854–1899.
Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. 2022. Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance* 77 (1): 5–47.
Garven, J.R. 2002. On the implications of the Internet for insurance markets and institutions. *Risk Management and Insurance Review* 5 (2): 105–116.
Gatzert, N., G. Schmitt-Hoermann, and H. Schmeiser. 2012. Optimal risk classification with an application to substandard annuities. *North American Actuarial Journal* 16 (4): 462–486.
Gemmo, I., M.J. Browne, and H. Gründl. 2019. *Privacy Concerns in Insurance Markets—Implication for market equilibria and social welfare*. ICIR Working Paper Series No. 25.
Gemmo, I., W. Mimra, and A. Sycheva. 2020. *A Franc less for a Pound more: (Price) discrimination and the value of privacy*. Unpublished Manuscript.
Gennaioli, N., R. La Porta, F. Lopez-de-Silanes, and A. Shleifer. 2022. Trust and insurance contracts. *The Review of Financial Studies* 35 (12): 5287–5333.
Geyer, A., D. Kremslehner, and A. Muermann. 2020. Asymmetric information in automobile insurance: Evidence from driving behavior. *Journal of Risk and Insurance* 87 (4): 969–995.
Gidaris, C. 2019. Surveillance capitalism, datafication, and unwaged labour: The rise of wearable fitness devices and interactive life insurance. *Surveillance and Society* 17 (1/2): 132–138.
Goldfarb, A., and C. Tucker. 2019. Digital economics. *Journal of Economic Literature* 57 (1): 3–43.

Guillen, M., J.P. Nielsen, M. Ayuso, and A.M. Pérez-Marín. 2019. The use of telematics devices to improve automobile insurance rates. *Risk Analysis: an Official Publication of the Society for Risk Analysis* 39 (3): 662–672.
Hartog, J., A. Ferrer-i-Carbonell, and N. Jonker. 2002. Linking measured risk aversion to individual characteristics. *Kyklos* 55 (1): 3–26.
Hassani, H., S. Unger, and C. Beneki. 2020. Big Data and actuarial science. *Big Data and Cognitive Computing* 4 (4): 40.
Ho, C.W.L., J. Ali, and K. Caals. 2020. Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance. *Bulletin of the World Health Organization* 98 (4): 263–269.
Hochschild, R. 1988. Biological age as a measure of risk. *Journal of the American Society of CLU and ChFC* 42 (5): 60–66.
Hoel, M., T. Iversen, T. Nilssen, and J. Vislie. 2006. Genetic testing in competitive insurance markets with repulsion from chance: A welfare analysis. *Journal of Health Economics* 25 (5): 847–860.
Holzapfel, J., R. Peter, and A. Richter. 2023. Mitigating moral hazard with usage-based insurance. *Journal of Risk and Insurance*. https://doi.org/10.1111/jori.12433.
Horvath, S. 2013. DNA methylation age of human tissues and cell types. *Genome Biology* 14 (10): 1–20.
Hoy, M. 1982. Categorizing risks in the insurance industry. *The Quarterly Journal of Economics* 97 (2): 321–336.
Hoy, M. 1984. The impact of imperfectly categorizing risks on income inequality and social welfare. *The Canadian Journal of Economics* 17 (3): 557–568.
Hoy, M. 2006. Risk classification and social welfare. *The Geneva Papers on Risk and Insurance: Issues and Practice* 31 (2): 245–269.
Hoy, M., and M. Durnin. 2012. *The potential economic impact of a ban on the use of genetic information for life and health insurance*. Ottawa: Office of the Privacy Commissioner of Canada.
Hoy, M., and M. Polborn. 2000. The value of genetic information in the life insurance market. *Journal of Public Economics* 78 (3): 235–252.
Hoy, M., and M. Ruse. 2005. Regulating Genetic Information in Insurance Markets. *Risk Management and Insurance Review* 8 (2): 211–237.
Hoy, M., and J. Witt. 2007. Welfare effects of banning genetic information in the life insurance market: The case of BRCA1/2 genes. *Journal of Risk and Insurance* 74 (3): 523–546.
Huang, H., M.A. Milevsky, and T.S. Salisbury. 2017. Retirement spending and biological age. *Journal of Economic Dynamics and Control* 84 (3): 58–76.
Infantino, M. 2022. Big Data analytics, InsurTech and consumer contracts: A European appraisal. *European Review of Private Law* 30 (4): 613–634.
Insilico Medicine. 2023. Deep biomarkers of human aging. Insilico Medicine. World Wide Web: http://www.aging.ai/.
Jeanningros, H., and L. McFall. 2020. The value of sharing: Branding and behaviour in a life and health insurance company. *Big Data and Society* 7 (2): 205395172095035.
Jin, Y., and S. Vasserman. 2021. *Buying data from consumers: The impact of monitoring programs in U.S. auto insurance*. Cambridge: NBER.
Kehr, F., T. Kowatsch, D. Wentzel, and E. Fleisch. 2015. Blissfully ignorant: The effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. *Information Systems Journal* 25 (6): 607–635.
Keller, A., and F. Transchel. 2016. Telematics: Connecting the dots. World Wide Web: https://www.swissre.com/Library/telematics-connecting-the-dots.html.
Kim, K.K., P. Sankar, M.D. Wilson, and S.C. Haynes. 2017. Factors affecting willingness to share electronic health data among California consumers. *BMC Medical Ethics* 18 (1): 1–10.
Kiviat, B. 2019. The moral limits of predictive practices: The case of credit-based insurance scores. *American Sociological Review* 84 (6): 1134–1158.
Koijen, R.S.J., and M. Yogo. 2023. *Financial economics of insurance*. Princeton: Princeton University Press.
Kotlikoff, L.J., and A. Spivak. 1981. The family as an incomplete annuities market. *Journal of Political Economy* 89 (2): 372–391.
Krippner, G.R. 2023. Unmasked: A history of the individualization of risk. *Sociological Theory* 41 (2): 83–104.
Krippner, G.R., and D. Hirschman. 2022. The person of the category: The pricing of risk and the politics of classification in insurance and credit. *Theory and Society* 51 (5): 685–727.

Lai, G.C., H. Nakamura, S. Yamamoto, and T. Yoneyama. 2021. Adverse retention: Strategic renewal of guaranteed renewable term life insurance policies. *Journal of Risk and Insurance* 88 (4): 1001–1022.

Lanfranchi, D., and L. Grassi. 2022. Examining insurance companies' use of technology for innovation. *The Geneva Papers on Risk and Insurance: Issues and Practice* 47 (3): 520–537.

Leverty, J.T., and J. Liu. 2019. Does technology adoption save regulatory compliance costs? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3517945.

Li, L. 2021. Opening up the black box: Technological transparency and prevention. *Journal Risk and Insurance* 88 (3): 665–693.

Li, L., and R. Peter. 2021. Should we do more when we know less? The effect of technology risk on optimal effort. *Journal of Risk and Insurance* 88 (3): 695–725.

Lindholm, M., R. Richman, A. Tsanakas, and M.V. Wüthrich. 2022. A multi-task network approach for calculating discrimination-free insurance prices. *European Actuarial Journal*. https://doi.org/10.1007/s13385-023-00367-z.

Liu, X. 2022. Picking lemons? Algorithm-aided human decisions in selection markets: Evidence from field experiments on insurance agents. PhD, University of Georgia.

Liu, X. 2023. Artificial intelligence and information production in selection markets: Experimental evidence from insurance intermediation.

Liukko, J. 2010. Genetic discrimination, insurance, and solidarity: An analysis of the argumentation for fair risk classification. *New Genetics and Society* 29 (4): 457–475.

Loi, M., C. Hauser, and M. Christen. 2022. Highway to (digital) surveillance: When are clients coerced to share their data with insurers? *Journal of Business Ethics* 175 (1): 7–19.

Lucivero, F. 2020. Big Data, big waste? A reflection on the environmental sustainability of big data initiatives. *Science and Engineering Ethics* 26 (2): 1009–1030.

Lünich, M., and C. Starke, 2021, Big data = big trouble for universal healthcare? The effects of individualized health insurance on solidarity. https://doi.org/10.31235/osf.io/3f2xs.

Mamoshina, P., K. Kochetov, E. Putin, F. Cortese, A. Aliper, W.-S. Lee, S.-M. Ahn, L. Uhn, N. Skjodt, O. Kovalchuk, M. Scheibye-Knudsen, and A. Zhavoronkov. 2018. Population specific biomarkers of human aging: A big data study using South Korean, Canadian, and Eastern European patient populations. *The Journals of Gerontology Series a, Biological Sciences and Medical Sciences* 73 (11): 1482–1490.

McFall, L. 2019. Personalizing solidarity? The role of self-tracking in health insurance pricing. *Economy and Society* 48 (1): 52–76.

McFall, L., and L. Moor. 2018. Who, or what, is InsurTech personalizing? Persons, prices and the historical classifications of risk. *Distinktion: Journal of Social Theory* 19 (2): 193–213.

McFall, L., G. Meyers, and I. van Hoyweghen. 2020. Editorial: The personalisation of insurance: Data, behaviour and innovation. *Big Data and Society* 7 (2): 205395172097370.

Meyers, G., and I. van Hoyweghen. 2018. Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. *Science as Culture* 27 (4): 413–438.

Meyers, G., and I. van Hoyweghen. 2020. 'Happy failures': Experimentation with behaviour-based personalisation in car insurance. *Big Data and Society* 7 (1): 205395172091465.

Milevsky, M.A. 2020a. Biological (and other) ages. In *Retirement income recipes in R*, ed. M.A. Milevsky, 259–279. Cham: Springer.

Milevsky, M.A. 2020b. Calibrating Gompertz in reverse: What is your longevity-risk-adjusted global age? *Insurance: Mathematics and Economics* 92: 147–161.

Milne, G.R., A.J. Rohm, and S. Bahl. 2004. Consumers' protection of online privacy and identity. *The Journal of Consumer Affairs* 38 (2): 217–232.

Montanera, D., A.N. Mishra, and T.S. Raghu. 2022. Mitigating risk selection in healthcare entitlement programs: A beneficiary-level competitive bidding approach. *Information Systems Research* 33 (4): 1221–1247.

Nayak, B., S.S. Bhattacharyya, and B. Krishnamoorthy. 2019a. Democratizing health insurance services; accelerating social inclusion through technology policy of health insurance firms. *Business Strategy and Development* 2 (3): 242–252.

Nayak, B., S.S. Bhattacharyya, and B. Krishnamoorthy. 2019b. Integrating wearable technology products and big data analytics in business strategy. *Journal of Systems and Information Technology* 21 (2): 255–275.

Nill, A., G. Laczniak, and P. Thistle. 2019. The use of genetic testing information in the insurance industry: An ethical and societal analysis of public policy options. *Journal of Business Ethics* 156 (1): 105–121.

Paefgen, J., E. Fleisch, L. Ackermann, T. Staake, J. Best, and L. Egli. 2013. Telematics strategy for automobile insurers. I-Lab Whitepaper, 1–31.

Page, M.J., J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, and D. Moher. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery* 88: 105906.

Palmer, D.E. 2006. Insurance, risk assessment and fairness: An ethical analysis. In *Insurance ethics for a more ethical world*. Research in ethical issues in organizations, 113–126. Bingley: Emerald (MCB UP).

Parente, S.T., D.S. Salkever, and J. DaVanzo. 2005. The role of consumer knowledge of insurance benefits in the demand for preventive health care among the elderly. *Health Economics* 14 (1): 25–38.

Peter, R., A. Richter, and P. Thistle. 2017. Endogenous information, adverse selection, and prevention: Implications for genetic testing policy. *Journal of Health Economics* 55: 95–107.

Pew Research Center. 2014. Public perceptions of privacy in the post-Snowden era. Pew Research Center.

Phelps, J., G. Nowak, and E. Ferrell. 2000. Privacy concerns and consumer willingness to provide personal information. *Journal of Public Policy and Marketing* 19 (1): 27–41.

Południak-Gierz, K., and P. Tereszkiewicz. 2023. Digitalization's big promise and peril: The personalization of insurance contracts and its legal consequences. In *Law and economics of the digital transformation*. Economic analysis of law in European legal scholarship, ed. K. Mathis, and A. Tor, 33–49. Cham: Springer.

Posey, L.L., and P.D. Thistle. 2021. Genetic testing and genetic discrimination: Public policy when insurance becomes "too expensive." *Journal of Health Economics* 77: 102441.

Pram, K. 2021. Disclosure, welfare and adverse selection. *Journal of Economic Theory* 197: 105327.

Putin, E., P. Mamoshina, A. Aliper, M. Korzinkin, A. Moskalev, A. Kolosov, A. Ostrovskiy, C. Cantor, J. Vijg, and A. Zhavoronkov. 2016. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging* 8 (5): 1021–1033.

Regner, T., and G. Riener. 2017. Privacy is precious: On the attempt to lift anonymity on the internet to increase revenue. *Journal of Economics and Management Strategy* 26 (2): 318–336.

Reimers, I., and B. Shiller. 2018. Welfare implications of proprietary data collection: An application to telematics in auto insurance. Available at SSRN 3125049.

Rohm, A.J., and G.R. Milne. 2004. Just what the doctor ordered. *Journal of Business Research* 57 (9): 1000–1011.

Rothschild, C. 2011. The efficiency of categorical discrimination in insurance markets. *Journal of Risk and Insurance* 78 (2): 267–285.

Rothschild, M., and J. Stiglitz. 1978. Equilibrium in competitive insurance markets: An essay on the Economics of imperfect information. In *Uncertainty in economics*, 257–280. Amsterdam: Elsevier.

Rothstein, M.A. 2015. Ethical issues in big data health research: Currents in contemporary bioethics. *The Journal of Law, Medicine and Ethics: A Journal of the American Society of Law, Medicine and Ethics* 43 (2): 425–429.

Rumson, A.G., and S.H. Hallett. 2019. Innovations in the use of data facilitating insurance as a resilience mechanism for coastal flood risk. *The Science of the Total Environment* 661: 598–612.

Saldamli, G., V. Reddy, K.S. Bojja, M.K. Gururaja, Y. Doddaveerappa, and L. Tawalbeh. 2020. Health care insurance fraud detection using blockchain. In *2020 Seventh international conference on software defined systems (SDS)*, 145–152. IEEE.

Samuel, G., F. Lucivero, and L. Somavilla. 2022. The environmental sustainability of digital technologies: Stakeholder practices and perspectives. *Sustainability* 14 (7): 3791.

Schubert, R., M. Brown, M. Gysler, and H.W. Brachinger. 1999. Financial decision-making: Are women really more risk averse? *American Economic Review* 89 (2): 381–385.

She, Z., T. Ayer, and D. Montanera. 2022. Can big data cure risk selection in healthcare capitation program? A game theoretical analysis. *Manufacturing and Service Operations Management* 24 (6): 3117–3134.

Siegelman, P. 2014. Information and equilibrium in insurance markets with Big Data. *Connecticut Insurance Law Journal* 21: 317.

Soyer, B. 2022. Use of big data analytics and sensor technology in consumer insurance context: Legal and Practical challenges. *The Cambridge Law Journal* 81 (1): 165–194.

Steinberg, E. 2022. Run for your life: The ethics of behavioral tracking in insurance. *Journal of Business Ethics* 179 (3): 665–682.

Strohmenger, R., and A. Wambach. 2000. Adverse selection and categorical discrimination in the health insurance markets: The effects of genetic tests. *Journal of Health Economics* 19 (2): 197–218.

Struminskaya, B., V. Toepoel, P. Lugtig, M. Haan, A. Luiten, and B. Schouten. 2020. Understanding willingness to share smartphone-sensor data. *Public Opinion Quarterly* 84 (3): 725–759.

Tanninen, M. 2020. Contested technology: Social scientific perspectives of behaviour-based insurance. *Big Data and Society* 7 (2): 205395172094253.

Tanninen, M., T.-K. Lehtonen, and M. Ruckenstein. 2022. Trouble with autonomy in behavioral insurance. *The British Journal of Sociology* 73 (4): 786–798.

Thiery, Y., and C. van Schoubroeck. 2006. Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance: Issues and Practice* 31 (2): 190–211.

Thomas, R.G. 2007. Some novel perspectives on risk classification. *The Geneva Papers on Risk and Insurance: Issues and Practice* 32 (1): 105–132.

Timms, P., J. Hillier, and C. Holland. 2022. Increase data sharing or die? An initial view for natural catastrophe insurance. *Geography* 107 (1): 26–37.

Verbelen, R., K. Antonio, and G. Claeskens. 2018. Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (applied Statistics)* 67 (5): 1275–1304.

Villeneuve, B. 2000. The consequences for a monopolistic insurance firm of evaluating risk better than customers: The adverse selection hypothesis reversed. *The Geneva Risk and Insurance Review* 25 (1): 65–79.

Villeneuve, B. 2005. Competition between insurers with superior information. *European Economic Review* 49 (2): 321–340.

Wiegard, R.-B., and M.H. Breitner. 2019. Smart services in healthcare: A risk–benefit-analysis of pay-as-you-live services from customer perspective in Germany. *Electron Markets* 29 (1): 107–123.

Wu, J.W., A. Yaqub, Y. Ma, W. Koudstaal, A. Hofman, M.A. Ikram, M. Ghanbari, and J. Goudsmit. 2021. Biological age in healthy elderly predicts aging-related diseases including dementia. *Scientific Reports* 11 (1): 1–10.

Xie, X., C. Lee, and M. Eling. 2019. Cyber insurance supply and performance: An analysis of the U.S. cyber insurance market. *SSRN Electronic Journal*.