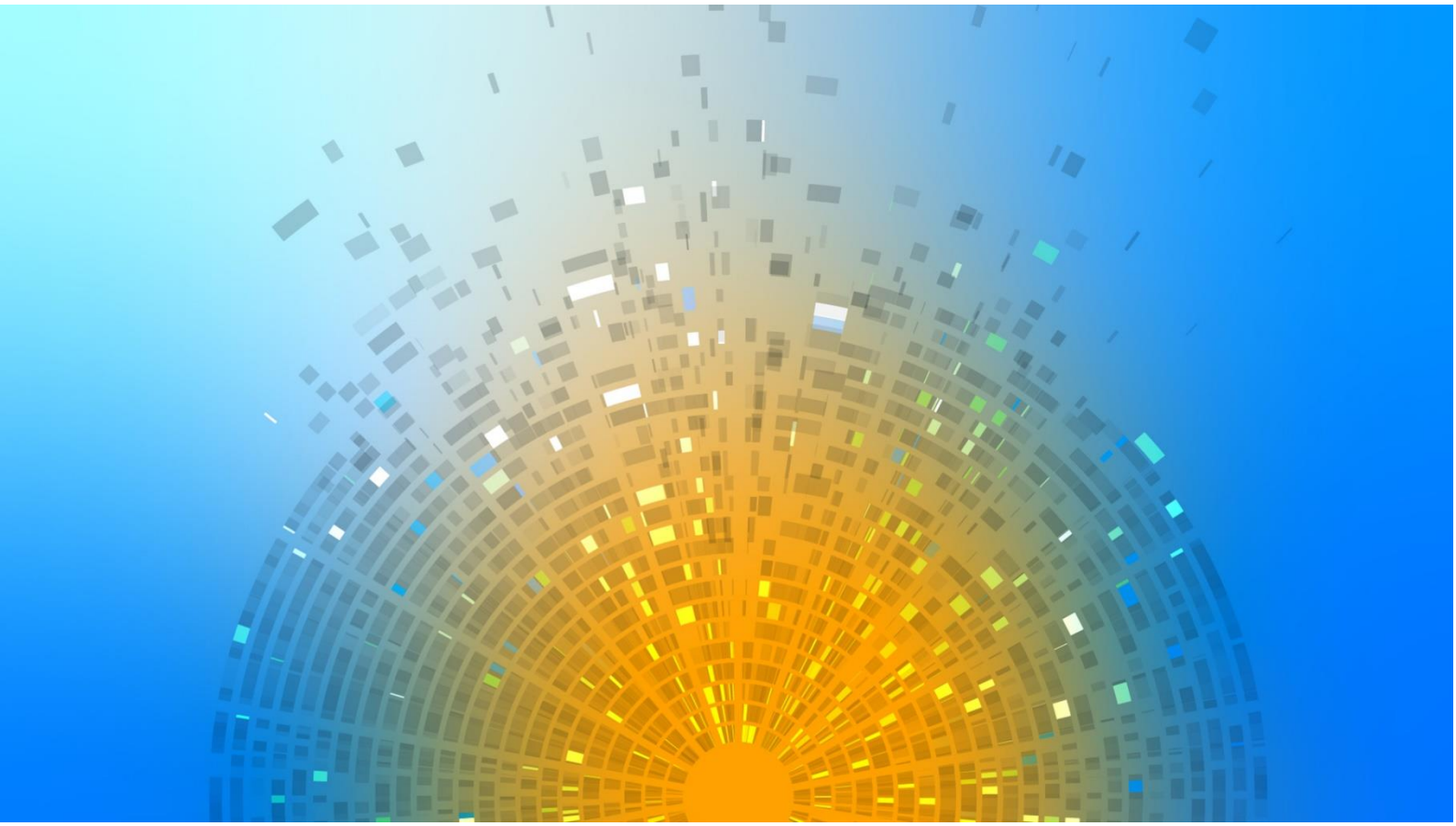
MILLIMAN REPORT

Potential data sources for life insurance AI modelling

April 2022

Bartosz Gaweda
Christoph Krischanitz
Remi Bellina
Jeff Anderson
Joe Long
Noriyuki Kogo
Saiki Justin Makino
Scott Chow



Milliman

Table of Contents

1. Executive Summary

We are living in exciting times. Every day more than 2.5 quintillion bytes of data are created, and that pace is only accelerating along with the growth of the Internet of Things (IoT). Vast columns of numbers are describing the world around us, starting with national statistics and financials compiled from our personal data, pictures of our neighbourhood, our activity on the internet, the style of driving a car, and even our inner world gathered by devices embedded under our skin. Such information will never be enough to perfectly forecast the future, assuming we are not living in a Matrix and there exists such a thing as free will. On the other hand, Big Data combined with the increased usage of Machine Learning algorithms, allows us to model the surrounding world much better than in the past and therefore to better understand underwriting risks.

Insurance companies have realized that maintaining the traditional business model is not enough anymore. They no longer have the monopoly on data and new startups called InsurTechs are appearing at a tremendous pace. New sales channels, data science teams, and others all aim to deliver the best possible customer service. The good news is that the life insurance industry is in a great position to deal with these challenges as their actuaries have been handling data and statistics for quite a while. Actuaries only need to gain access to these new data sources and apply innovative solutions.

Is it so simple then? The fact that actuaries have an easier start here than people from outside the industry is of course beneficial; however, they still struggle with difficulties and need to learn new skills. The data is everywhere around us, but it needs to be properly gathered. Then it is necessary to deal with data quality, which never is perfect. What is even more important is not all the data can be used because of ethical issues and regulations, which varies around the globe. Being stricter by nature, it is the responsibility of insurance companies to be a leader for taking care of such ethical considerations.

At the end of the day the usage of external data can improve customer service by providing personalized approaches and better prices thanks to lower costs for insurance companies. Moreover, due to risk prevention models using better data, it is possible to improve the health of customers by offering what is beneficial both to them and to insurance companies. On the next several pages, we will explore in detail what kind of data can be used to achieve these goals and what kind of limitations may be encountered in the process.

2. Living in the Cyber World

The insurance industry faces a big shift from product focus to the client-centricity model. For this reason, we will start with one specific client. Let us call him Bob Snow.

It is 7 am. Bob wakes up in the morning. He played video games long into the night so his smartwatch registered that he slept only for five hours. The scoring of his sleep is not so high either. He may have general problems with sleeping or, possibly, he had a few beers while playing video games. Regardless, he does not feel sleepy, his smartwatch woke him up in a light sleep phase. We can see that his pulse increased as he stood up. Automatic sensors detect the movement in the room and open the blinds (what confirms the information from the smartwatch). In the meantime, a coffee machine is preparing a double espresso (of course the information is integrated with his smartphone and goes directly to the cloud). He goes into the bathroom where he uses his Bluetooth-enabled toothbrush (which rewards good brushing habits[1]). Already, his smartwatch has registered 134 steps. Then Bob chooses in his app one of the proposed healthy breakfasts (he also eats a few pieces of his favourite chocolate, but somehow forgets to mention it in the app!). While drinking coffee he creates a shopping list on the smartphone.

---

[1] The role of wearables in private medical insurance, available at https://www.milliman.com/-/media/milliman/pdfs/articles/the_role_of_wearables_in_private_medical_insurance.ashx

‘Don’t forget to activate your new running shoes!’ – says Bob’s wife when he goes out for a morning jog. Sensors gather all the steps during the run. The smartwatch proves the data—this day Bob did his best and ran 15 km (nearly twice the height of Mount Everest). It appears that, for his age (he is in his fifties) he is in a good shape. Have we already mentioned that Bob is a tech fan and wears smart shorts, smart underwear, and a smart belt[2]?

We do not think we need to convince anyone that the times of Artificial Intelligence (AI) have already come. We read everywhere about vast amounts of data gathered every day. Google data centre is one of the world’s largest buildings with more than 2 million square feet (about twice the area of Chicago's Millennium Park) of usable space.[3] Forecasts predict that there will be around three billion 5G subscriptions worldwide by the year 2024, which speeds up data transmission by up to twenty times compared to 4G/LTE,[4] what only accelerates the process of data gathering. At the same time, it takes a few seconds for an AI to learn huge amounts of information via quantum computers[5] and current projections suggest that devices and smart sensors enabled by the IoT will generate at least 500 zettabytes of data by 2020.[6] If that was not enough, Generali creates an AI model which translates the cry of a new-born[7] and Disney’s robot performs amazing acrobatics.[8] One can say that the world is quite shocking already.

Every day modern technologies go deeper and deeper into the insurance industry. Deep learning is used for cancer screening or to help children with disabilities.[9] Tesla is entering the insurance industry for their partially autonomous cars.[10]

Considering all the above, it is sometimes hard for life insurance incumbents to catch up when history for some of them reaches back to the 18th century. The opportunity to exploit this situation has been detected by many startups (or InsurTechs). Furthermore, many insurance players feel the threat of the entry of data giants such as Google or Amazon into the market.

Currently, the life insurance market is even more difficult than before. With significant decrease of margins and the low interest rate environment (or even negative interest rates in some countries), companies’ actuaries need to be more careful than ever about long-term assumptions such as mortality or retention rates as the impact on future projections is discounted at a lower discount rate. In the year 2020, all insurance companies needed to face the new pandemic reality. Substantial changes can be seen not only in the access to new data and algorithms, but also in the attitude of clients who have become much more demanding.

There is not one single response in the market to the above issues, but some common points can be observed, such as shifting to digital sales channels, improving customer relations, or cooperating with other service providers. Another obvious response would be to test the possibilities using AI models and understanding clients better by using external data.

[2] Impact of Wearables and the Internet of Things. Available at https://www.actuaries.org.uk/practice-areas/health-and-care/disbanded-research-working-parties/impact-wearables-and-internet-things

[3] How Big Is A Google Data Center? (akibia.com). Available at https://www.akibia.com/how-big-is-a-google-data-center/

[4] 5G - Statistics & Facts | Statista. Available at https://www.statista.com/topics/3447/5g/

[5] 25+ Impressive Quantum Computing Statistics [Updated for 2021] (seedscientific.com). Available at https://seedscientific.com/quantum-computing-statistics/

[6] Life insurance and the Internet of Thinking - Accenture Insurance Blog. Available at https://insuranceblog.accenture.com/life-insurance-and-the-internet-of-thinking

[7] Assicurazioni Generali 2021 | EFMA Innovation In Insurance. Available at https://innovationininsurance.efma.com/assicurazioni-generali-2021

[8] Disney’s Stunt Robots Could Change How Hollywood Makes Action Movies | Movies Insider - YouTube. Available at https://www.youtube.com/watch?v=nZ950ywJy0M

[9] 10 Wonderful Examples Of Using Artificial Intelligence (AI) For Good (forbes.com). Available at https://www.forbes.com/sites/bernardmarr/2020/06/22/10-wonderful-examples-of-using-artificial-intelligence-ai-for-good/?sh=451cd5ca2f95

[10] Tesla Insurance: Everything You Need to Know (caranddriver.com). Available at https://www.caranddriver.com/car-insurance/a35434414/insurance-tesla/

In this paper we will describe what external data is available for the life insurance industry, how it can be used and, importantly, how it should not be used. However, please note that we focus here on data. For more details about AI models, we encourage you to look at Milliman paper, 'The use of Artificial Intelligence and Data Analytics in Life Insurance' (to be issued in Q4 2021). Moreover, please keep in mind that potential variables described in this paper are not exhaustive, and it is possible to continue searching for new ideas, which we strongly encourage. Their usability for an insurance company portfolio must be assessed, if applied.

3. Data Around Us

3.1 What is Big Data?

Just to make sure that we are clear, we present our definition of Big Data. These words are used very often so it is important to know what we are talking about.

Big Data has become the latest buzzword. In today's connected world, there is an effervescence of data from multiple sources.

We often speak about Big Data to denote a particular or innovative context. Let's clarify things. First, it is important to note that the term 'Big Data' is difficult to grasp. Even today arriving at a precise and clear definition is challenging because the term covers many distinct uses.

However, it is possible to agree on the following main principles:

- This term refers to data sets whose volume and complexity are such that traditional data processing software is not able to manage and process them.
- In general, data processing is expected to be completed within an acceptable timeframe.
- It is also worth noting that the notion of Big Data is synonymous with diversity in data. Thus, structured data (information managed in databases) as well as unstructured data (information that is not organized and does not correspond to a model or a predetermined format) are both in scope.
- In the insurance world, the term 'Big Data' is often used even if several petabytes are not involved, as long as the context breaks with the existing IT processes for example.

In the literature, Big Data is thus characterized by the three Vs:

- A large **volume** of data
- A wide **variety** of data types
- The **velocity** at which the data must be processed and analysed

A fourth V can sometimes be added, related to data **veracity** and data quality issues.

The sources of data can be the web, the IoT (even more with the arrival of 5G), etc. Almost everything today can be a generator of data, but data without analysis is nothing, as is analysis without conclusions and actions. Not everything is good to capture or integrate into models. What if information about something that might seem innocuous is correlated with, say, the mortality rate of the population? Without even talking about causality, machine learning models are quite capable of highlighting relevant variables in the modelling of a phenomenon, even if the variables are numerous. When this is accomplished, companies can make better informed decisions.

The notion of Big Data, as promising as it is, necessarily comes with a set of components that allows organizations to use data in a practical way with the right IT infrastructure—the storage systems and servers designed for Big Data, which can be in-house or outsourced to a cloud. To store all the data, organizations set up suitable data storage infrastructures (traditional data warehouses, data lakes, etc.). One of the most representative technologies is the Hadoop system. The Apache Hadoop project develops open-source software for distributed computing.

Understandably, efforts in Big Data and analysis require specific skills in IT (Hadoop, Spark, etc.) as well as in Data Science (machine learning, data quality, etc.).

It is worth noting that large data sets are not in and of themselves Big Data. Even if an insurance company has millions of customers, the personal data, the policy data, and the claims data together cannot be classified as Big Data in the original sense. Very often such large data sets can be handled easily with traditional or even advanced statistics. It really takes more Vs to become Big Data.

3.2 Data Science methods

Like the words Big Data, we see that Data Science is often misinterpreted.

3.2.1 WHAT IS DATA SCIENCE?

Data Science is a field of applied mathematics and statistics that provides useful information based on large amounts of complex or large amounts of data. The concept itself is not new. The history of artificial neural networks (ANN) began with Warren McCulloch and Walter Pitts (1943) who created a computational model for neural networks based on algorithms called threshold logic.[11] Generalized Linear Models, which are a generalization of linear regressions, have been used by actuaries for decades. The only thing which changed is the fact that we have more data, more computational power, and easier access to resources and ready-to-use, open-source packages. The most popular algorithms in the insurance industry are:

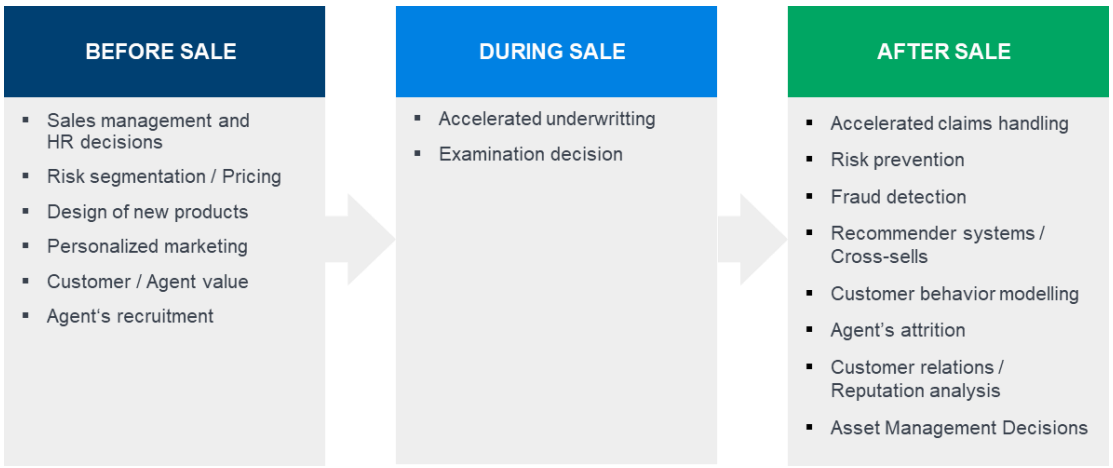
- **Supervised methods** such as: Neural Network, Deep Learning and Tree Based Models (Random Forest, Gradient Boosting Machine (GBM), eXtreme Gradient Boosting Model (XGBoost), Light GBM)
- **Unsupervised methods**[12] such as: K-Means Clustering, Hierarchical Clustering, K Nearest Neighbors (KNN), Principal Component Analysis (PCA)

One common mistake is referring to robots helping to automate processes (RPA, Robotic Process Automation) as machine learning. Such robots make sure that repetitive work is done automatically; the difference is that they do not 'learn' with each iteration the way machine learning algorithms do. If a robot repeats its work for 100 years, it will not become any smarter.

Data Science uses techniques such as machine learning and artificial intelligence to extract meaningful information and to predict future patterns and behaviours. Describing the technical details of these methods is out of scope of this paper, but you can find more of them in the Milliman research paper, 'The use of Artificial Intelligence and Data Analytics in Life Insurance' (to be issued in Q4 2021). For now, it will be enough to say that machine learning is nothing scary and overly complicated. In case you would like to understand the magic behind it right now, many good and publicly available trainings are at your fingertips.

3.3 Use cases

FIGURE 1: POTENTIAL BENEFITS OF USING DATA SCIENCE AT EACH MARKETING PHASE



[11] History of artificial neural networks – Wikipedia. Available at https://en.wikipedia.org/wiki/History_of_artificial_neural_networks

[12] In unsupervised learning, the algorithms use unstructured data set to reach a certain conclusion (like creating bins, called clusters, which are separating the whole data base in the best possible way). Only input variables are given. The opposite is true for supervised methods where we have access to the target variable.

With knowledge of Big Data and Data Science methods we can now focus on their applications. The truth is that only our imagination limits the possible models that can be created, but we gather below the ones which are most popular, bring the most added value, or have the most potential in their future development.

3.3.1 HUMAN BEHAVIOR

Usually, it is the job of agents or customer service to understand what we can expect from specific clients. It is hard to perform uniform actions across the whole portfolio because different policyholders behave in diverse ways. Thanks to machine learning algorithms it is possible to understand clients' behaviours better and to estimate what is the best response to them. In the subsections below we will describe the following possible use cases: lapsation, funds switching, agents' attrition, HR solutions, recommendation systems, and cross selling.

It should be clear that human behaviour cannot be explained perfectly by data analysis. Even with tons of data, not every aspect of a human's decisions is predictable. And if there is something like a 'free will' the predictability of data analysis on individuals is very limited. However, there are a lot of typical patterns in human behaviour which are not explored so far, and machine learning algorithms offer good technology to learn about and make use of them.

Retention, recruitment, and funds switching models

One key issue for insurers is how to retain policyholders. For long-term insurance, lapse can have significant impacts on future cash flows. Earlier literature mentions how lapse rates of some life insurance products such as variable annuities were influenced by the financial market. In conventional dynamic lapse modelling, lapse rates are assumed to depend only on interest rates. Our objective in this section is to expand from conventional literature to more generalized lapse modelling beyond the dependency on interest rates. A deeper understanding of policyholder behaviour in combination with focused action can lead to improvements in policy retention. In the beginning, actuaries may think—'Is this really necessary? Maybe a simple assumption that 5% of policyholders will lapse yearly is enough?' Questions like these are understandable but retention models open a way not only to better understand risks but also to provide a more customized approach to policyholders.

With regards to explanatory variables for lapse models, a variety of external and internal features can be considered. External features can include indices like GDP and unemployment rate, while internal features can include variables like age of policyholder, region (culture), or type of product. In addition, new variables can be created by using interactions of multiple variables, for example, by region and income. For more external and internal features, please refer to the tables below:

FIGURE 2: POSSIBLE EXTERNAL FEATURES FOR LAPSE MODELING[13]

External Features	Additional Information
Macro-economic variables	
GDP	
Buyer confidence	
Inflation	
House price development	
Economic growth	
Return on stock market	
Unemployment	
Equity market volatility	
Interest rate volatility	If interest rates change significantly, it can have large impacts on customer behaviour.
Exchange rates	
Time variables	
Seasonal effects	
Other variables	
Reference market rate	What are other companies offering?

Source: Michorius, 2011

[13] MSc_CZ_Michorius.pdf (utwente.nl) http://essay.utwente.nl/61317/1/MSc_CZ_Michorius.pdf

FIGURE 3: POSSIBLE INTERNAL FEATURES FOR LAPSE MODELING[14]

Internal Features	Additional Information
Contract Specific Variables	
Type of product	Depending on whether you are building a model for term life, whole life, unit-linked, pension, etc., the explanatory variables you use will be different.
Age of contract	Policy year of contract.
Lifetime of contract	Insured period of contract.
Premium frequency	Annual, semiannual, monthly, or single, etc.
Premium size	
Value of insurance	Type of coverage, insured amount, etc.
Surrender charge	
Company-specific variables	
Distribution channel	Agent, broker, bank, internet, or direct mail, etc.
Negative publicity	Negative publicity could lead to mistrust by policyholders and increase lapse rate.
Crediting rate	
Policyholder variables	
Age of policyholder	
Gender	
Widowed	
Marital status	
Postal code (region)	
Income	
Mortality rate	

Source: Michorius, 2011

FIGURE 4: OTHER VARIABLES THAT COULD BE INCLUDED IN A LAPSE MODEL

Other Features	Additional Information
Frequency of contacting policyholders	If contacting policyholders too often or too little, this could have negative effects on lapse rate.
Sentiment of policyholder contacts with insurance company	Sentiment analysis can be done to analyse how positive or negative interactions with the customers have been, which can affect lapse rate.
Policyholder health using wearables	If a policyholder has poor health, which can be detected using wearables, we expect the tendency of lapse to be lower.

With a model in place and a deep understanding of the policyholders with elevated risk of lapse, different strategies can be used to retain these policyholders more effectively. For example, a rate change can be done in a segment where a competitor is offering better rates, more contact can be made with policyholders that get contacted too little, or other products can be offered to a policyholder that went through a life change (marriage, bought a house, etc.). These are just a few strategies that could be used to retain policyholders and, as the understanding of policyholders improves, more strategies can be created.

Agent attrition and recruitment

It is worth mentioning that retention models apply not only to a client of insurance companies but to agents as well. Using a similar approach, we may want to estimate which agents are most likely to quit their job (taking their clients to the competition, or just leaving them without proper care, which can lead to the client's lapse). Such models (both client and agent retention) can be used both during the contract (policy or employment) and at the beginning of it (while pricing products and HR processes, respectively).

[14] Ibid.

It is possible to model not only the probability that a given agent will quit his job quickly but also what profit he will probably bring to the company. Bear in mind we can help HR departments customize the recruitment process and to answer questions such as:

1. Should we look for new employees in big cities or villages? The question is correlated also to estimating Agents' Lifetime Value and distribution of sales.
2. What is their desired level of education and work experience?

At the same time, we need to be aware that all such models need to be properly constructed and tested to avoid any kind of discrimination (for example by sex, age, or race).

Fund switching

Moreover, using a similar set of variables you (or your actuary) can model the behaviour of clients in terms of funds switching between more risky assets and safer options like bonds. Depending on the profile of the client we can expect different reactions to the market situation and, based on the results, better estimate future charges on the portfolio.

Sales recommender and cross selling systems

If a person calls an agent and wants to buy life insurance, should we also propose an accidental death rider? What about investment products, or cancer, or hospital attendance riders? What kind of product and what sum insured should we offer? We can base this decision on characteristics of the buyer (age, address, credit scoring, friends, occupation, number of kids), but also on the history of earlier purchases and conversation (How many times have we spoken already? Does a client have any insurance? In which company?). Based on the historical databases about purchases of the whole portfolio, we can train models to answer questions that are important in the sales process. This idea is not theoretical. Plenty of insurance players build such models internally and/or acquire InsurTechs to optimize their sales.[15]

3.3.2 UNDERWRITING SUPPORT

Accelerated Underwriting

One of the main problems of traditional life underwriting is its timing. When a client needs to wait even one month for the underwriting decision (mostly due to the need of lab results and medical exams), she can simply make up her mind or go to competitors. Because of that, the insurance industry seeks solutions to accelerate the process and make underwriting decisions much faster. To achieve this goal, a few options are currently used:

1. Decreasing the number of questions[16]
2. Putting behavioural economics into action[17]
3. Using image/video/audio recognition to create new data points (described further in sections Internal Data and External Data/Face recognition)
4. Using data from smart devices (described further in section External Data/IoT)
5. Using external data such as health history, drug prescription history, vehicle driving history, and credit scoring (described further in External Data chapter)

Studies show that using external data instead of standard examinations is not only sufficient but can even improve model performance.[18] To avoid regulatory issues, one solution is to classify clients to the best segment using machine learning models and to pass the rest of the clients to the standard path.[19] Thanks to such an

[15] Life Insurance Sales Recommender System: December 2020 (milliman.com). Available at https://www.milliman.com/-/media/milliman/pdfs/2020-articles/articles/12-16-20-life_insurance_sales_recommender_system-v1.ashx

[16] Smart underwriting: 4 + 2 health questions suffice for in-depth risk assessment | Munich Re Topics Online. Available at https://www.munichre.com/topics-online/en/life-and-health/smart-underwriting.html

[17] Improving Accelerated Underwriting Results by Putting Behavioral Economics Into Action | Gen Re. Available at https://www.genre.com/knowledge/blog/improving-accelerated-underwriting-results-by-putting-behavioral-economics-into-action-en.html

[18] Stratifying Mortality Risk Using Physical Activity as Measured by Wearable Sensors | Munich Re US Life. Available at https://www.munichre.com/us-life/en/perspectives/wearables/Stratifying-mortality-risk-using-physical-activity-as-measured-by-wearable-sensors.html

[19] Life Insurance Application Assessment Prediction | by Weichen Lu | Data Science is life | Medium. Available at https://medium.com/time-to-fish/life-insurance-application-assessment-prediction-484910062678

approach, no adverse actions are taken with the usage of AI and external data. The underwriting department's employees can now focus on more challenging tasks while simple ones are carried out by machines.

Last but not least, AI models are responsible also for chat and voice bots which answer the most standard client questions. There is an anecdote told by the representative of one of the InsurTechs on a great online seminar,[20] when a client said that he was able to find his policy online, fill out all required positions with the help of a chat bot, buy a policy and pay for it, all while waiting on the line to get the answer from an agent at one of the traditional life insurance companies.

Decision who should be sent for an examination

Usually, clients of a certain age who ask for a product with a sum insured above some specified limit are asked to get an additional medical examination. Such tests are not only costly, but as described in the earlier section, can be time-consuming. Thanks to machine learning models, it is possible to assess in which cases it is worth sending clients for additional examination and in which it does not. Consequently, the insurance company reduces costs, gains more clients, and improves customer relationships.

3.3.3 CLAIMS HANDLING SUPPORT

Accelerated claims handling

In analogy to accelerated underwriting, employees of claims handling departments can obtain help from machine learning models. After initial classification, most simple claims cases can be solved and paid, even within seconds (with the help of chat/voice bot), while more difficult cases are passed to employees. Again, the insurance company can reduce costs and improve customer relationships.

Fraud detection

As a result of circumstances described above, some claims (and some sold policies) can be marked as a high probability of fraud. Thanks to the usage of external data, models can recommend which claims should be treated with utmost care by the claims handling departments' employees.[21]

3.3.4 PRICING

Pricing may be treated as the most crucial element of life insurance business. Research done by a multinational insurance company shows that the machine learning approach can outperform traditional underwriting by better subdividing lives by risk, such that the healthiest risk pool had a 6% reduction in deaths after 15 years compared with traditional underwriting.[22] Machine learning can also be used to derive the assumptions for use in traditional pricing models. One example is in the forecasting of mortality; (Hainaut, 2018)[23] used neural networks to forecast mortality rates in France between 2001 to 2014 generated by a model trained on general population mortality rates between 1946 and 2000. The model showed that the mortality forecasts based on neural network models had a higher predictive power when compared with a range of Lee-Carter models, the industry standard approach. Such an approach was applied to U.K. mortality data in (Richman, 2020b).[24] [25]

[20] Pricing and Longterm Underwriting Methods Using AI | Insurtech Insights - YouTube. Available at https://www.youtube.com/watch?v=7SjEObizHLU

[21] Investigating Life Insurance Fraud and Abuse (rgare.com). Available at https://www.rgare.com/knowledge-center/media/research/investigating-life-insurance-fraud-and-abuse

[22] LifeScore Labs_Med360.pdf (hubspotusercontent40.net). Available at https://f.hubspotusercontent40.net/hubfs/5627392/LifeScore%20Labs_Med360.pdf

[23] (PDF) A NEURAL-NETWORK ANALYZER FOR MORTALITY FORECAST (researchgate.net). Available at https://www.researchgate.net/publication/322344759_A_NEURAL-NETWORK_ANALYZER_FOR_MORTALITY_FORECAST#:~:text=Many%20researchers%2C%20including%20%28Hainaut%202018%29%20in%20his%20paper%2C,detecting%20latent%20time%20processes%20while%20directly%20predicting%20mortality

[24] AI in actuarial science – a review of recent advances – part 2 | Annals of Actuarial Science | Cambridge Core. Available at https://www.cambridge.org/core/journals/annals-of-actuarial-science/article/abs/ai-in-actuarial-science-a-review-of-recent-advances-part-2/C35A295A1F3ECC3013EA4D953706694A

[25] Artificial Intelligence: The ethical use of AI in the life insurance sector (milliman.com). Available at https://uk.milliman.com/-/media/milliman/pdfs/2020-articles/articles/11-24-20_ethics-ai_20201117.ashx

3.3.5 ASSET MANAGEMENT

Another example of the application of AI can be found in Asset Management. The key applications of AI for investment management that were identified in the 2019 CFA Report[26] were:

1. **To generate new data:** To extend the data that can be analysed to support investment decisions, by using NLP (natural language processing), computer vision and speech recognition to process text, images, and audio data, respectively.
2. **To support systematic investment strategies:** Leveraging machine learning to improve the algorithms used by quantitative investment managers in their processes.
3. **To support active fund management decision making:** To get new investment insights for active fund managers by using AI techniques to process Big Data, including the use of alternative and unstructured data (e.g., satellite images, earnings conference call recordings and transcripts, social media postings, consumer credit and debit card data, and e-commerce transactions).

3.3.6 RISK MANAGEMENT

Most life insurance contracts can be valued in a reasonably straightforward way using discounted cash flow models that employ a range of software applications, from Microsoft Excel to proprietary asset liability modelling software. However, in certain cases, alternative techniques to estimate the output of a more complex cash flow projection model may be appropriate, and these could potentially use AI. Such techniques may be considered in cases where:

1. Stochastic simulations considering many potential future scenarios are needed (e.g., for with-profit contracts and other contracts with options and guarantees).
2. Nested simulations are needed (e.g., in the calculation of the Solvency II capital requirements) that may necessitate, or otherwise increase, the number of model runs and simulations required.
3. Decision makers in insurers require an understanding of the solvency position more often than monthly or quarterly valuations would allow (i.e., daily solvency monitoring) considering changes in the economic environment.

Proxy modelling has been in use in the U.K. and Europe for several years, principally to support the calculation of the solvency capital requirement (SCR) in Solvency II internal models. The overall goal of proxy modelling is to create a relatively simple function to estimate the output of a more complex cash flow projection model, such that a loss function (the fit of the model to the data) is minimized. [27] As explained before, such simple models describing more sophisticated cash flows can be done using machine learning models.

[26] https://www.cfainstitute.org/-/media/documents/survey/AI-Pioneers-in-Investment-Management.ashx

[27] Artificial Intelligence: The ethical use of AI in the life insurance sector (milliman.com). Available at https://uk.milliman.com/-/media/milliman/pdfs/2020-articles/articles/11-24-20_ethics-ai_20201117.ashx

3.4 Data travel from being Big to being Useful

FIGURE 5: BIG DATA ANALYTICS PROCESS

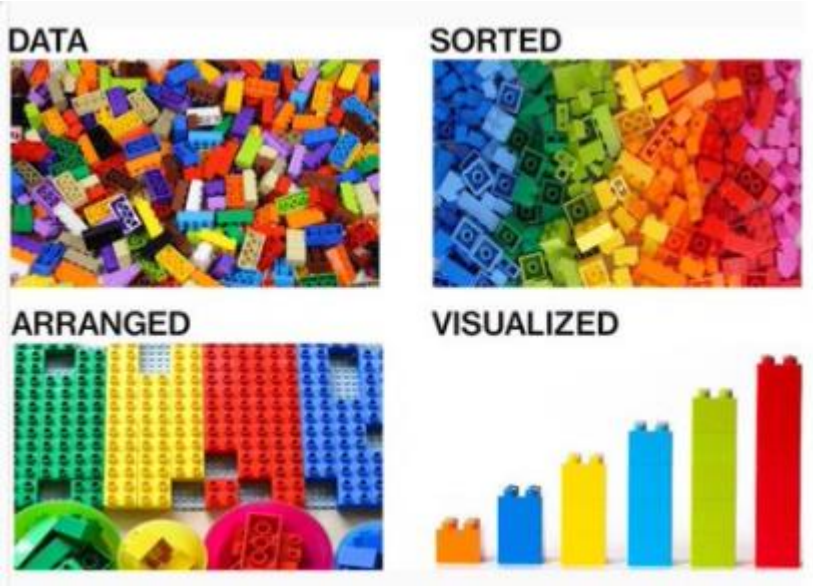
Big Data takes a while before it goes from being Big to becoming Useful. We could say that it starts with Data Collection but that would be too short-sighted. The first stage—and maybe the most important one—is to determine what data should be gathered. Then, find out what aspects of this data are not available and start a process to gather it.

Then dirty work comes in. '…76% of data scientists view data preparation as the least enjoyable part of their work.' At the same time 'data scientists spend around 80% of their time on preparing and managing data for analysis.'[28] After pre-processing, we can proceed with the modelling phase which is usually assessed as the most interesting one. By the way, considering that 76% of data scientists do not like the part of their work that takes 80% of their time, is quite surprising considering that this job is still called 'the sexiest job of the 21st century.'

Lastly, we have a ready-to-use model that has been evaluated so we know it performs in line with our expectations. Still, a few steps are ahead. We need to deploy the results and to monitor them to make sure everything works as we intended. The model predicts the outcome; however, the result must be checked further before taking to the client (in the case of web aggregators) or to company employees, to recommend, for example, if any claims should be treated with greater caution because of a high probability of fraud. This is an inevitable, challenging part of the process. At the end of the day, this is what often characterizes the work done by actuaries—no model is good before it is used in practice. And figuring that out is what actuaries do best.

[28] Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says (forbes.com). Available at https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=47d945086f63

FIGURE 6: BIG DATA CAN BE MODELLED MANY WAYS TO SEEK INNOVATIVE SOLUTIONS



MOST COMMON PROBLEMS

Very often creating sophisticated AI models is the easiest part of an AI project. The problems occur both at the beginning with the quality of input data and at the end with deploying results (to agents, web aggregators, claim handlers, customer service consultants):

- When we speak about external data for underwriting, results need to be obtained a maximum of a few days after application, if not in real time, to enable prediction of model.
- When we speak about risk prevention, we need to have a real-time connection with a client's devices.
- When we speak about customer behaviour models, we must watch the behaviour of markets in real time or download the most recent unemployment statistics in each region to be able to react fast enough.
- Especially when we speak about external data, pre-processing and validation may be needed. Many variables can be unstructured (like pictures, videos, or text) and before we use them, they must be transferred to a proper format.
- Furthermore, we need to be extra careful to ensure the data is correct so additional reconciliation and/or using different sources for comparison is a good practice.

3.5 Data gathering

The data discussed in this paper may come directly from the policyholder (e.g., via a form or indirectly from scanned documents) or from external sources. These external data are generally available on the Internet and can be linked to the client or future client by means of a key: age, a specific social situation, or the postal address, which is often one of the main keys. It is then a question of linking external information according to the individual's place of residence, using the geolocation of the address to integrate it into a larger data field.

No matter what data you are looking for, it is often necessary to retrieve it through an API (Application Programming Interface). Let us take the example of geographic data manipulation using Google APIs.

FIGURE 7: RETRIEVING GPS COORDINATES WITH THE CURL FUNCTION IN CMD WINDOWS USING THE GOOGLE API:

```
curl "https://maps.googleapis.com/maps/api/streetview/metadata?location=arc+triomphe+paris+france,&key=%KEY%"
{
   "copyright" : "© Google",
   "date" : "2020-03",
   "location" : {
      "lat" : 48.87426641926022,
      "lng" : 2.294809965904432
   },
   "status" : "OK"
}
```

FIGURE 8: RETRIEVAL OF A SATELLITE IMAGE WITH A REQUEST MADE IN PYTHON USING THE GOOGLE API:

```
address='Arc de Triomphe Paris France'

params = {'center':address,'maptype':'satellite','size':'600x600','key':API_key}

map_response = requests.get('https://maps.googleapis.com/maps/api/staticmap',params=params)
```

It is also possible to use APIs from other data providers such as OpenStreetMap. For example, we can extract the list of hospitals around a given address. We notice that APIs are particularly useful to retrieve, massively for all individuals, complementary information that allows us to enrich the models or to lighten the information requests. But what exactly do we mean by APIs?

An API serves as a gateway, an intermediary between two computer systems, which can then understand each other. Each API manages a specific task and regulates access to data. This access is done in real time: if the data source is updated, the request to retrieve it will also be updated. One of the main advantages of an API is that you do not need to know the internal workings of the software at the source. The growing use of APIs is having a significant impact in all industries, including insurance: They increase the ability to access diverse data sources and innovate faster.

The term REST API is also often used and is the most popular style for creating web APIs: It is a standard way for web services to send and receive data. REST stands for Representational State Transfer and determines the API specifications through a set of rules. The request accesses a specific URL and sends a request before receiving a proper response. HTTP requests are typically sent (e.g., GET and POST). The response can be in different formats (e.g., JSON). Such APIs are easy to deploy and maintain.

Increasingly widespread, also in connection with the rise of open data and open-source initiatives, APIs are making a significant contribution to the diversification of insurance data and its ease of access. They also provide a concrete and authorized framework for data recovery, sometimes in opposition to the notion of web scraping.

Collecting valuable data for analysis is the foundation of any data science process. When data comes from an external online source, one method is indeed web scraping. This technique consists of extracting data from a website using a robot, or rather a script (e.g., in R or Python). In general, companies are aware that this process is possible, and try to protect their data by IT barriers (extensive use of JavaScript and captchas for example), or more simply by legal notices. In any case, there is today a real effervescence around data, allowing us to enrich and innovate, especially in the field of life insurance.

3.5.1 NUDGING

To speak about nudging it may be best to start with an example:

There was a study looking at a policy review form used by a US vehicle insurance company on which drivers had to declare on a form how many miles they had driven in the previous year. The idea being, the more miles you declare, the more your car is used, and therefore your premium should be higher. At the end of this form there was an honesty statement saying, 'I promise that everything I have written in the above is true and correct.' And so, what the insurance company did was simple: They moved the honesty statement from the bottom of the form to the top.

They divided 13,000 drivers receiving this policy review form into two groups and randomly allocated them to place the honesty statement either at the bottom or the top of the form.

The findings? Drivers who signed after, at the end of the form, declared 30,095 km, on average. Those who signed before declared 42,000 km, on average. That difference was worth $48 per insurance premium for the insurance company, over $500 million for the whole group. And again, it was a small change with a disproportional effect.[29]

Nudging means giving small incentives to clients, which can help them make buying decisions or prevent fraud. From the perspective of data gathering, such incentives can help improve data quality and prevent fraudulent answers because studies show people are more committed to telling the truth. Other examples of nudging would be to create web articles or blogs which answer questions often asked by clients or even marketing actions. To assess the success rate of such nudging techniques statistical tests could be used.

3.6 Data quality

Data we gather and collect is rarely of the required quality. Usually, there are some missing values, problems with data structure, formatting, or availability. Old data may not even be stored in an electronic version, but nowadays that is not much of a problem anymore. Although it takes a lot of time, assessing the quality of data and correcting it to the extent possible is necessary. At the end of the day, that is likely the best data we could have and using it is better than doing nothing.

The good news is that everyone has this problem and thanks to that many solutions are easily available. Most machine learning algorithms (like tree-based models) can treat missing values as a separate level of the variable or even impute the best estimate of its value based on other variables. However, such solutions need to be treated with the highest caution as missing values can be caused, not only by random events, but also by process changes and/or time effects.

Using external variables can help fill missing gaps as well, but even more importantly, they can help validate the accuracy of the data. For example, using data from external providers can help you accelerate underwriting and ask less questions. However, if quality of the information matters more than speed, you may both ask the client for the answer and time validate it with an external source. It is not only possible to improve the quality this way, but also to identify client profiles that are less eager to be honest, for example.

[29] How Nudges Reduce Insurance Fraud: Small Changes Make a Difference - FRISS. Available at https://www.friss.com/blog/how-nudges-reduce-insurance-fraud-small-changes-make-a-difference/

3.7 Data administration

If you have been working in an insurance company long enough, you have faced the most common problems already. You know that because of the speed of changes in information technology, systems are updated and changed constantly, or even are being exchanged with other ones once in a few years. Data infrastructure is not getting simpler, rather the opposite—new subsystems are being created and the integration between them is not always perfect. The fact that old data still is stored in paper does not make it easier. At the same time, existing processes need to be updated along with the systems. On the other side, with the creation of new processes, management information systems, and reporting requirements etc., new systems are being created that close the circle. All these changes must be made in line with still enlarging lists of data protection and security requirements (especially when we consider that 'Cyber-attacks occur 2,244 times per day,' as counted by University of Maryland[30]). Additionally, you might have encountered situations where, due to staff rotations, some datasets are left completely unattended and finding the right path for data takes too much time.

The same situation happens when we consider external data. Not only do we need to store and handle it properly, we also need to make sure these processes are done in line with regulations. On the other hand, we need to keep in mind that we are not the owner of external data and at some point, it can stop being available. We also are not in control of the frequency of new updates. There are no guarantees. Very often data administration issues will create hurdles and delays for new and even ongoing AI projects.

But external data provides a lot of advantages. Apart from added insight into current clients, it gives us the ability to fill missing values in old records in case we lost some of the information or simply did not decide to store it in the past. And, because these problems are faced by all (insurance) companies, those who are dealing best with this issue have a big advantage over the others. The key is to be aware that regular changes and adaptions will happen in future and have processes in place which are adaptive and flexible, and to anticipate future developments as early as possible. Be attentive and active.

3.8 Appropriateness of data

It is not an easy task to test external variables for such models as mortality. Normally to perform such an assessment we need a big historical data set of internal and external variables and a target variable (in this case the target variable would be the indicator of a death). To do it for long term life insurance products, we need years of exposure, and we may not even know what the external variables looked like at the moment of underwriting. Because of that, it is not easy to assess their impact on past events.

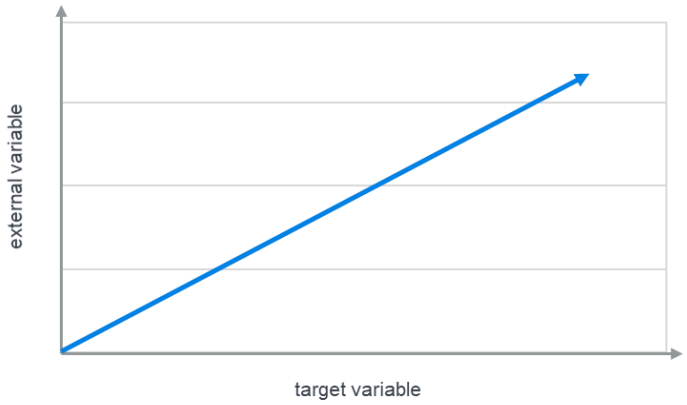
It is easier when we speak, for example, about one-year term life insurance—we can gather external data for all policies issued within some period of time and, after one year, test for the significance of external variables. We could even assume that during this one-year period, external variables did not change significantly and look only at the year passed without the need to wait an entire year (with proper prior testing if such assumption is reasonable). For the sake of simplicity, let us assume then that we are talking about a term insurance contract for one year.

To test whether new variables will help to improve the performance of an existing model, a few steps should be considered:

- First, it is good to perform univariate analysis, meaning, to draw a chart, where on the x axis we have a new external variable and on the y axis we have target variable (for example lapse ratio). The result can be a monotonously increasing curve. We may see that with the increase of our external variable, the target variable increases as well. Such a chart may show some correlation, but not necessary show causality.

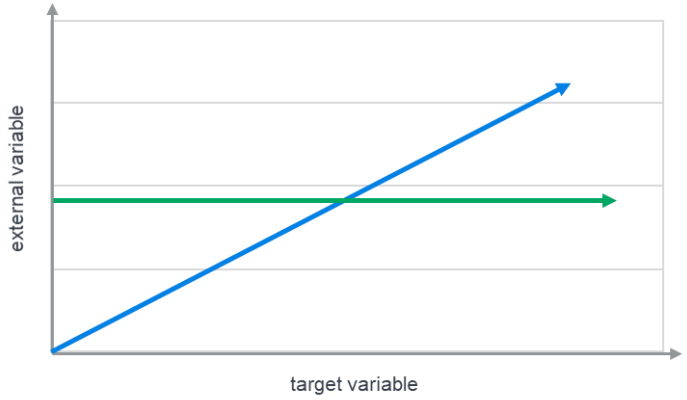
[30] According to a study by the University of Maryland, a cyber-attack occurs once every 39 seconds. Cyber-attack statistics derived from the study indicate that the most attempted username was 'admin.' The most attempted password was the same word, and it was used in 43% of all hacking attempts. (University of Maryland) 30 Cyber Security Statistics & FAQs for 2021 - 99firms. Available at https://99firms.com/blog/cyber-security-statistics/#:~:text=Cyber%20security%20attacks%20statistics%20show%20that%2032-33%25%20of,data%20invasions%2C%20however%2C%20involve%20some%20form%20of%20hacking

FIGURE 9: CORRELATION BETWEEN TARGET AND EXTERNAL VARIABLES



- It may be the case that the new variable is highly correlated with some internal variable we had already (for example, we added the height of the person nominated in meters, but originally had it in centimetres). First, we need to determine whether the 'old' model was explaining this increasing pattern already. To accomplish this, we must add the old model prediction to the chart (the best option would be to do it on a test set not used for training of the old model, to avoid impacting potential overfitting[31]). If we learn that the old model was blind for this pattern, we may assume that the new data gives additional insight.

FIGURE 10: COMBINING OLD MODEL DATA WITH NEW VARIABLES TO GAIN ADDITIONAL INSIGHTS



- After first visualization, we see that first candidates provide a good new variable. Now we can go to the modelling phase (it is also possible to start here and skip the two earlier steps, especially if we have plenty of variables). We assume that the old model was created well and know its performance (measured by some error measure on the test set or by using a cross-validation scheme). Depending on the type of model, a few possibilities are in store:
  - In the case of GLM (generalized linear models) it may be best to use stepwise linear regression and/or the ANOVA method to assess which variables are adding the most value. However, even in times of fast computers, if we test many variables at the same time, such processes may be very time consuming. After creating the model with the best selection, we can compare the performance with the old model. Penalization terms can also be applied to the size of coefficients when training a GLM to aid in feature selection. One such method is a Least Absolute Shrinkage and Selection Operator (LASSO) regression, which is much faster than stepwise selection methods. Using a LASSO regression, one can include all variables in the model, whereby non-significant variables will have their coefficients shrunk to zero.

[31] By overfitting we call the situation, when models fit very well (too well) to training set data and is not performing well on the test set.

- In the case of tree-based models, we can add all variables at once. Thanks to more sophisticated algorithms such as GBM (gradient boosting machine) or XGB (extreme gradient boosting), a considerable number of variables will not cause a problem. We move at once to compare performance. Without additional efforts we can look at the importance of each variable to determine which variables are bringing the most value to the model. Then we can use methods called 'Interpretable Machine Learning' to 'unbox the black box' (for example, partial dependence plots, Shapley Value; for more please refer to another paper created by Milliman specialists called 'Interpretable Machine Learning for Insurance'[32]). Then, even if the regulator of your country is not comfortable (yet) with tree-based methods, we can use the results of the above analysis in more traditional modelling.
- A few other methods, for example, Principal Component Analysis (PCA) can help reduce the dimensionality of data before using supervised learning algorithms.

No one method is best for all cases. The right solution may be to try a few methods and take what works best—especially when each dataset is different.

3.9 Internal data

The insurance industry is such a promising branch for using predictive modelling because it has possessed massive amounts of statistical data for many years. The following are typical variables which are in the hands of companies already:

- Basic profile: age, sex, name, ID number, address of residence (country, street and number, ZIP code, region, county), email address, telephone number, bank account
- Product dependent: sicknesses in family, regular medications, pre-existing medical conditions, occupation, education level, yearly salary, marital status, children, being a smoker or an excessive drinker, being a marijuana smoker or addicted to drugs, weight, height, criminal record, habits, and hobbies
- Results of lab results and/or physician examinations
- Policy characteristics: amount of premium, sum insured, riders, premium frequency, type of policy, type of payment, commission
- Sales channel, selling unit, administration unit, internal adviser, other insurance policies, main contact person within insurer

It is common that insurance companies gather plenty of other data but do not use it (perhaps because of privacy reasons or lack of time or resources). Examples include:

- Conversion rates from aggregators or agents
- History of policies and claims from other products and/or lines of business (like P&C or health policies). Information given during underwriting for these (e.g., model of car, value of home, buying travel insurance including extreme sports coverage)
- History of chats/calls/video calls with customer service
- History of emails / post communications
- Historical addresses
- Historical actions (e.g., switching between funds in investment products)

Companies such as Lemonade create thousands of new data points that describe their clients based on numerous touchpoints including natural language processing of written communication.[33] Also, it is technically possible to use voice recognition and to analyse historical calls and video chats, if allowed by regulations.

Thanks to using conversion rates, it is possible to estimate price elasticities for different segments of clients and, as a result, to optimize margins and volume sales using dedicated software. It should be mentioned that price elasticity models are currently under scrutiny by regulators in the UK due to a potential lack of fairness. Please check current regulations in your country before applying such a solution.

[32] Interpretable Machine Learning for Insurance (milliman.com) available at https://www.milliman.com/-/media/milliman/pdfs/2021-articles/4-2-21-interpretable-machine-learning.ashx

[33] Lemonade is 5 Years Old! Here's Our Strategy and Results to Date – YouTube available at https://www.youtube.com/watch?v=j7Q8SyuHWc0

It is worth also considering the information about postal addresses. We have separate addresses for almost every policy. We have thousands of postal codes. We will describe reasons for this later in section 4.1.3 GIS. For now, we note that geography is not heavily used in life insurance models, although it can provide a lot of information about policyholders' behaviour.

Currently, one of the greatest issues life insurance companies are facing is the small number of touchpoints. The first touchpoint is the moment of signing a policy. The second can be the moment of claim settlement. Depending on the type of policy, it may even be the contact with the beneficiary, not with the policyholder anymore. During the long duration of a contract the client may change his behaviour dozens of times. He could change his address or even country. He could start smoking, or his health status could change drastically.

Because of these facts, we see a few crucial trends in the business:

- Attempts to increase the number of touchpoints by wellness programs, risk prevention, gamification, and cross-selling opportunities in perfect timing or by connecting insurance with other services (like connecting to the bank industry through PSD2 regulation,[34] or creating ecosystems where, apart from buying insurance, you can handle other services like paying bills).
- Extreme care about anticipation of touchpoints along with personalized and often digitalized customer service to deliver the highest level of service.
- Usage of external data to fill missing gaps in knowledge about customer behaviour.

4. External Data Sources

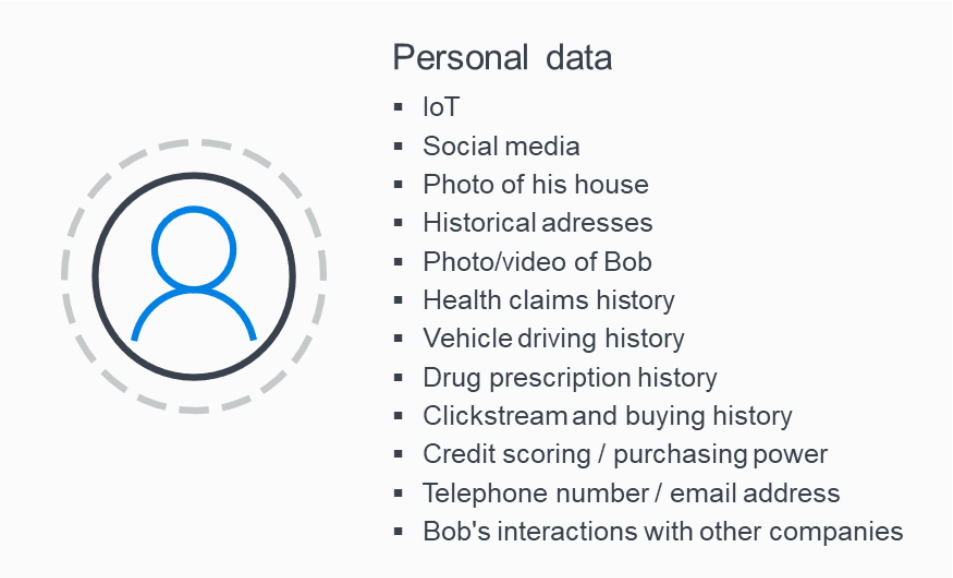
We know there is plenty of data out there. We know what goals we would like to achieve and what kind of models we need to create to achieve these goals. So, let us go more into practice. What variables are available and where can we find them?

4.1 Types of data

4.1.1 CONNECTED ON A PERSONAL LEVEL

In the Introduction section, you learned about Bob. It is possible to know quite a lot about him as described in the Internal Data section. Now we will focus more on what else we can find out about him using external data sources. After this, we will try to answer questions such as: Where do we get the data? What do we gain from this? Can we do it?

FIGURE 11: TYPES OF PERSONAL DATA COMMONLY AVAILABLE



[34] Payment Services Directive 2 - all you need to know (jpmorgan.com). Available at https://www.jpmorgan.com/europe/merchant-services/insights/PSD2-all-you-need-to-know

Internet of Things

The story in the Introduction stopped at about 9 am when Bob came back home after running. Now we just stress what happens next:

- Bob does not like to spend too much time driving; he **exceeded speed limits by about 150% for most of his way**.
- **GPS location** shows that Bob works **50 km from home**, **between coal-fired power plants**, where he works as an office worker (we know this from the company's internal data). The average **pollution there is 10 times higher** than at his home. He spends there **on average 9 hours a day**. Is this work really safe?
- After coming back home he checks his **blood pressure** and **body temperature** using his medical kit.
- During the day sensors gathered a few more data points, like:
  - **Galvanic Skin Response** (GSR, which falls under the umbrella term of electrodermal activity, or EDA), refers to changes in sweat gland activity that are reflective of the intensity of our emotional state, otherwise known as emotional arousal.[35]
  - **How many times has Bob coughed today?**
  - **What was his respiration rate?**
  - **What was the level of sugar in his blood** (using device embedded under the skin)?

As you can guess, we have not mentioned many other sensors like **brain activity measures, ingestible** (broadband-enabled digital tools that we actually eat, for example, 'smart' pills that use wireless technology to help check internal reactions to medications[36]), and many others.

At the end of the day, Bob falls asleep. Thousands of variables are stored in his smartphone. They will be gathered for the next 365 days. Some of the above data points are gathered for more than 25% of the US population up to the age of 49[37] (the percentage is even higher for younger groups with whom insurance companies attempt to engage the most). The average UK home has 10 internet-enabled devices according to a study from Aviva.[38]

We can classify the above examples into categories:

- Wrist-borne wearables
- Medical devices
- Smart clothing and shoes
- Other wearables, including jewellery
- Embeddable
- Ingestible

44% of consumers are likely to consider connected insurance services to help them become and stay healthier.

[35] What is GSR (galvanic skin response) and how does it work? - iMotions. Available at https://imotions.com/blog/gsr/

[36] Ingestibles, Wearables and Embeddables | Federal Communications Commission (fcc.gov). Available at https://www.fcc.gov/general/ingestibles-wearables-and-embeddables

[37] 21% of Americans use a smart watch or fitness tracker | Pew Research Center. Available at https://www.pewresearch.org/fact-tank/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/

[38] Tech Nation: number of internet-connected devices grows to 10 per home - Aviva plc. Available at https://www.aviva.com/newsroom/news-releases/2020/01/tech-nation-number-of-internet-connected-devices-grows-to-10-per-home/

How can I use it?

Do you think such data can improve the performance of models described in section 3.3? What about fraud detection? What about product pricing? Or risk prevention? Can they replace standard variables like blood results or physician examinations to improve the speed of processing (accelerated underwriting)? Do you think people who join such a program will serve as a portfolio for people with healthy lifestyles and be used the insurance company to benefit from positive selection or the opposite? And even more importantly, will such data improve the health level of the clients? Or maybe data privacy concerns will not allow its use? We will comment on the last question in more detail in the separate section for regulations. For now, according to our knowledge, there are already companies which successfully use such data.

FIGURE 12: EXAMPLES OF CURRENT USE OF WEARABLES IN INSURANCE BASED ON ANOTHER MILLIMAN REPORT:

Aditya Birla Health Discounts for policyholders who record a specified number of steps using an activity tracker or attend gym sessions or have a health assessment.	**The Vitality Programme** Vitality members earn points and achieve a higher Vitality status when they undertake activities that are assumed to impact on health status. Higher Vitality statuses unlock higher rewards for benefits such as gym, travel and other discounts.	**AXA** Offers a free Withings Pulse fitness tracker. Participants receive discounts of over $100 on their insurance policies, as well as discounts off any Withings product purchases when they complete a certain number of steps.
Oscar Rewards customers who track their fitness data gift cards when they reach their step goals.	**United Healthcare** Rewards users with healthcare credits for reaching daily fitness goals.	**Qantas Assure** Policyholders receive Qantas frequent flyer points if they lead more active lifestyles.
Aetna Monitors daily activity and provides assistance in achieving personalised health goals. The app also provides recommendations, nudges and rewards.	**Esurance** SavorBand devices are offered which can capture information on food consumed, including recipes, cooking tips, and purchasing discounts along with other data.	**Beam Technologies** Uses Bluetooth-enabled toothbrushes to reward good brushing habits with discounted insurance premiums and other rewards.

Moral hazard and fraud risk

Apart from regulation and data privacy issues, the main threat for wearables can be the risk of fraud. It may not be easy to fool the device which is embedded under your skin and measures sugar in your blood, but buying a device which will artificially increase the number of steps measured by your smartwatch is quite simple. You can always turn off or remove your smartwatch if you plan to behave in an 'unhealthy' way, similar to a turning off a tracker in your smartphone when you plan to drive your car faster (a big problem for telematics).

Insurance companies will need to find a way to confirm data, for example, by using a few various sources (do you remember the example with Bob's smart shoes and smartwatch?).

What can we gain from this?

To check whether this question is worth pursuing, we can use the results of publicly available reports. As described in the report from analysis conducted by Munich Re,[39] 'There is robust evidence that physical activity as measured by steps per day can effectively segment mortality risk even after controlling for age, gender, smoking status and various health indicators.' Another analysis conducted by Milliman specialists resulted in the conclusion that 'wearables may encourage members to increase their activity levels, however, the implementation should be considered as a part of comprehensive wellness offering.'

[39] Stratifying Mortality Risk Using Physical Activity as Measured by Wearable Sensors | Munich Re US Life. Available at https://www.munichre.com/us-life/en/perspectives/wearables/Stratifying-mortality-risk-using-physical-activity-as-measured-by-wearable-sensors.html

Where can I get the data?

At some point you (or your actuary) need to perform tests by yourself. The architecture of the test will depend on the model you want to create, but first you need to obtain data which obviously is not lying on the street. No one will send it to the insurer for free. The answer is to engage clients and convince them they will benefit from sharing such information. Such benefits can vary significantly from discounts on initial premium, to profit sharing during the life of their policy. The fact that clients will obtain smart devices can be treated as a benefit as well. Results from a survey done by Accenture show, 'around 65 percent of millennials globally would consider a connected life insurance product.' [40]

Furthermore, insurance companies are rarely popular and trusted by customers. Here, at last, there is an opportunity to be 'on the same side as clients' because both have common interests. Of course, insurance customers want to be healthy as long as possible and that is exactly what insurance companies want them to be. The consumers will likely be happy if they receive data-driven recommendations about their health and how to improve it. Finally, many people like technological innovations and information. Recommendations based on AI might encourage them to buy the policy. Access to the information can be beneficial, not only to the insurer but to the person insured as well. A project run by AXA for property protection showed that clients were eager to share more information on receiving feedback about the levels of risk in their home along with risk prevention tips.[41] Such cooperation also allows insurance companies to obtain more touchpoints with clients and improve customer relationships.

Most probably you will need to cooperate with the device provider to assure dataflow to your systems. The list of potential companies can be found in another Milliman paper.

Social Media

We start with the assumption that insurance companies are not restricted either by law or by ethics. The goal of such a thought experiment, which is obviously not realistic, is to find out what kind of data is available out there and how insurance companies could use it. Restrictions are complex and depend on the use case. Please look at section 4.2 to find out more about the status of regulations in this area worldwide.

We will take Facebook as an example. Please note that some countries have banned or temporarily limited access to Facebook including Mainland China, Iran, Syria, and North Korea.[42] In a few of these countries, Facebook is not so popular (for example Japan). However, a similar approach could be used to analyse data from Messenger, LinkedIn, Twitter, WhatsApp, Telegram, Instagram, TikTok, Reddit,[43] VKontakte (VK),[44] LINE,[45] and other social media websites.

Facebook

As at the end of the year 2020, more than 70% of US and Canada adult citizens were using Facebook. Already, you probably know very well what kind of data is being shared there, but just to gather it in one place, we create a list of possible variables below:

- Number of friends
- Indicator that his friend was caught in fraud (understanding fraudulent nets of people)[46]
- Date of account creation (was it a few years ago or more recent?)
- Level of activity

[40] Connected wellness opens a myriad of opportunities for life insurers - Accenture Insurance Blog. Available at https://insuranceblog.accenture.com/connected-wellness-opens-a-myriad-of-opportunities-for-life-insurers

[41] Give Data Back: sharing to better protect you and your home |... (axa.com). Available at https://www.axa.com/en/magazine/give-data-back

[42] Censorship of Facebook - Wikipedia retrieved on December 15, 2021 from https://en.wikipedia.org/wiki/Censorship_of_Facebook#:~:text=Many%20countries%20have%20banned%20or%20temporarily%20limited%20access,been%20restricted%20in%20various%20ways%20in%20other%20countries

[43] Having as many users as Twitter, **Reddit** is one of the greatest sources of UGC (User Generated Content) in the world. Reddit also provides public APIs that can be used for a variety of purposes such as data collection, automatic commenting bots, or even to assist in subreddit moderation.

[44] **VK** is a Russian social media platform geared toward Russians and other Eastern European users.

[45] LINE is a freeware app for instant communication, popular in Japan

[46] Detecting Life Insurance Fraud – YouTube. Available at https://www.youtube.com/watch?v=iFKNihULKys

- Belonging to groups (are these parental groups or extreme sky diving groups?)
- Likelihood of travel, music preferences
- Extremely sensitive information like political views, religious beliefs, race, gender, or sexual orientation
- Age, marital status, education, current and earlier occupation (does it match the company's internal data?)
- Family, relatives, children, age of children
- Address (does it match the company's internal data?), earlier places of living
- Profile picture / other pictures / videos
  - Drinking alcohol
  - Smoking
  - Driving a bike without a helmet
  - Rock climbing without protection
  - Driving fast cars
  - Or, is he supposed to be unable to work (he claims that he is) and then uploads a picture as a Salsa instructor from this weekend?
- Pictures / posts on his family / friends / secondary accounts
  - Have they just published a photo from a ski trip where our client was actively present, while claiming that his leg is broken?
  - Or, they shared info that he broke his leg on Friday, while he bought travel insurance on Saturday and claims that the leg was broken on Sunday?
- Content of posts, style of writing.

Based on the above list it is not very surprising that 'Facebook activity data alone could indicate your psychological make-up more accurately than your friends, your family—better even than your partner, given enough info […] as proved by experts from The University of Cambridge and Stanford University back in 2015, in which they examined the Facebook profiles of more than 86,000 participants, and then matched their on-platform data against their psychological profiles, which those users had submitted through a personality test app.'[47]

As concluded in the paper, 'Social Media Networking Data Analysis in Life Insurance Underwriting'[48] written in Taiwan in 2015, many of the above variables were already given by the client during the underwriting process and, because of that, can only be used for validation purposes. However, with information such as credit scoring, moral[49] and morale[50] hazards, the situation becomes different and, if insurance companies are allowed to gather this data, it may help to properly underwrite the risk of the individual.

Many investigators report that navigating an insured individual's online social media page is one of the first things they do when looking into potentially fraudulent claims, according to a report from Boston-based research firm Celent in 2011. The Internet is full of hints on how to use social media in insurance fraud detection, like seeking inconsistent details, finding confessions, and/or checking secondary and friends' accounts.[51] [52] [53] [54] [55]

[47] What Does Facebook Know About You Really? | Social Media Today. Available at https://www.socialmediatoday.com/news/what-does-facebook-know-about-you-really/546502/

[48] Social Media Networking Data Analysis in Life Insurance Underwriting, IJAIEM-2015-04-05-8.pdf. Available at https://www.ijaiem.org/Volume4Issue4/IJAIEM-2015-04-05-8.pdf

[49] Moral hazard refers to behavioral changes that might occur and increase the risk of loss when a person knows that insurance will provide coverage.

[50] Morale hazard describes the same kind of behavioral change that might occur when a person knows insurance will cover them, but in this case, it's subconscious.

[51] https://riskandinsurance.com/social-media-tool-claims-investigation/

[52] 8 tools for using social media to fight insurance fraud | PropertyCasualty360. Available at https://www.propertycasualty360.com/2017/01/27/8-tools-for-using-social-media-to-fight-insurance/

[53] Social media helps insurers manage underwriting, claims and risks in real-time | Business Insurance. Available at https://www.businessinsurance.com/article/20130630/News07/306309992

[54] 3 Ways to Use Social Media in Insurance Fraud Investigations | i-Sight. Available at https://i-sight.com/resources/3-ways-to-use-social-media-in-insurance-fraud-investigations/

[55] 4 Tips to Successfully Use Social Media Discovery to Prove Insurance Fraud (claimsjournal.com). Available at https://www.claimsjournal.com/news/national/2018/10/02/286988.htm

Where can I get the data?

Even apart from regulatory and ethical issues there are plenty of other factors which can make using social media in models difficult:

- Many people use social media; many do not. That means insurance companies can gain more detailed insight into portfolios of those who do, but still be blind in this area to those who do not.
- People are increasingly cautious about the value of their data. For this reason, many restrict publicly available content to be accessible only to their friends. On the other hand, people seem more eager to publicly share their professional image on LinkedIn, however, such attitudes are restricted to people from specific industries.
- Remember Bob Snow? If we find 55 such profiles on Facebook, the list may include Bobby Snows, Snow Bobbies, Robert Snows, etc.… We can try to restrict our search to some city, some age, telephone number, email, or other information based on internal variables to decrease the number of results, but in the end it may be hard to determine which of these is the right person. There are at least two remedies for this issue:
  - First, you (or your actuary) may use a few various sources to confirm your findings. Only if a few of them confirm the same information may it be recognized as valuable.
  - Second, you may convince a client to connect his social media account to his client's insurance profile. Again, as in the section about wearables, the client will need to see a benefit to himself before he makes such a decision.
- Because of lack of created API connections, we would need to use web scraping solutions to get access to the data (please refer to section 3.5). Please note that Facebook explicitly forbids any kind of web scraping without written permission from Facebook. In response to the public outcry following the Cambridge Analytical scandal, Facebook implemented dramatic access restrictions on its APIs in April 2020.[56,57]

```
# Notice: Collection of data on Facebook through automated means is
# prohibited unless you have express written permission from Facebook
# and may only be conducted for the limited purpose contained in said
# permission.
# See: http://www.facebook.com/apps/site_scraping_tos_terms.php[58]
```

Taking the above into consideration, we do not recommend trying to automatically obtain data from Facebook. Even though it appears that restrictions for other social networking websites are less strict, we need to keep in mind that they may eventually follow Facebook's approach. No one wants to create any kind of solution which can be stopped at any moment. However, taking the above into consideration, we are not saying you should abandon any attempts to obtain social media data. However, if that is your intent, make sure that all interested parties (regulators, insurance companies, clients, social networking websites) do not have any objections to the applied solution. For more information about regulatory and ethical issues please refer to section 4.2.

Photo of a house

In some countries such as the US, data availability regarding client credit scoring and purchasing power is better. However, in some countries, it is hard to assess client details from the perspective of his wealth. Many sources suggest that such information can greatly improve the performance of models (what we can interpret from life insurance, for example, as a level of access to healthcare or to healthy food, but also can be correlated with a level of education or insurance awareness). We need to consider that even if the client supplies us with a level of yearly salary, that data alone may not be completely accurate.

[56] 5 Things You Need to Know Before Scraping Data From Facebook | Octoparse. Available at https://www.octoparse.com/blog/5-things-you-need-to-know-before-scraping-data-from-facebook

[57] Facebook forces Admiral to pull plan to price car insurance based on posts | Car insurance | The Guardian. Available at https://www.theguardian.com/money/2016/nov/02/facebook-admiral-car-insurance-privacy-data

[58] https://www.facebook.com/robots.txt https://www.facebook.com/robots.txt

For these reasons, external data sources may help to verify the information. In the US, you can refer to the following sections about credit scoring and purchasing power, but it does not limit your options. On the website called Zillow[59] (please note that it may not be the only possibility for this purpose), you can check the value of the house based on an address, but also to create variables such as:

- How many bathrooms does it have?
- Does is it have a garage?
- What kind of heating is used?
- And others

Another study conducted by the Massachusetts Institute of Technology (MIT) proved that by creating deep learning techniques for picture recognition of house images accessed by Google Street View, you can create powerful variables to improve motor pricing.[60] Further research needs to be done to assess the value of such a solution in the life insurance industry; however, it may be worth the effort. Life insurance business differs in many ways from P&C, including longer contracts during which a person can change their address.

Historical addresses

It is quite common for people to move in their lifetime. The number of moves varies by country; however, individuals in the US tend to move more frequently than those in most other countries—estimated to be, on average, 11 times in their lifetime. Having information on where people have lived can be useful for various analyses. For example, such data can be used to study general migration patterns of a population to measure how various census demographics change over time. With accurate address histories on individuals, we can create longitudinal datasets where we can attach information about the environments in which people have lived using Geographic Information System (GIS) mapping. This information can be incorporated into predictive models to investigate the correlation between mortality and moving patterns, among other things. For example, an analysis of the distance a person lives from a hospital could be analysed to see potential impacts on a person's mortality.

Comprehensive databases that track an individual's address change history across countries are currently not available to the public. However, in the US, multiple sources collect and aggregate address change information on individuals moving within the US. To name a few, this includes sources such as change-of-address requests submitted to the postal service, vehicle registration departments, and utility companies. This information can be purchased from credit bureaus if you have a permission to use this information. The major US companies that can provide this information are Equifax, Experian, and TransUnion.

Additionally, this information can be bought from various aggregators and has historically been used for marketing purposes. However, the data has many other uses that might be of interest to life insurers. Not only can it be useful in studying mortality or morbidity of individuals, but also to update policyholders' contact information as they move to ensure mail is sent to the correct address. It can also be used to identify fraud by flagging people who have unusual moving patters.

Photo / video recognition

The idea of using face recognition is not new in the life industry. Pictures can help to access biometric details of a person (like sex or age) to accelerate underwriting. However, recent studies show that this technology can do much more. In the analysis called 'AI Enabled Next Generation LTC and Life Insurance Underwriting Using Facial Score Model,'[61] we recognize that face recognition models can identify visual symptoms, such as:

- Facial morphology – abnormal characteristics
- Facial asymmetry
- Abnormal skin colour
- Yellowish face or eye colour

[59] How Much is My House Worth? Check Your Zestimate | Zillow. Available at https://www.zillow.com/how-much-is-my-home-worth/

[60] [1904.05270] Google Street View image of a house predicts car accident risk of its resident (arxiv.org). Available at https://arxiv.org/abs/1904.05270

[61] AI Enabled Next Generation LTC and Life Insurance Underwiring Using Facial Score Model. Available at https://insurancedatascience.org/downloads/London2021/Session_4b/Shrinivas_Shikhare.pdf

- Abnormal eye movement or disturbances in facial expressions
- Abnormal muscular response
- Abnormal head pose

To create initial diagnoses for health issues like:

- Down Syndrome
- Facial paralysis
- ADHD
- Pain
- And many others

The authors of the above-mentioned analysis shared also interesting ideas about what can follow face recognition, including Biosensors and audio, video, and image-based analysis. Once again, we note that such analysis needs to be conducted with the utmost caution to avoid any kind of discrimination.

Health claims and drug prescription history

When an individual wishes to purchase a new life insurance product, they typically go through a medical examination. Scheduling such a procedure is expensive and takes time, and might deter individuals from purchasing an insurance contract. However, the process can be streamlined if the underwriters can pull medical and prescription drug histories in an automated way.

In the US, information from past underwriting applications can be pulled from the Medical Information Bureau (MIB). This data enables potential insurers to determine whether applicants omitted any health information in the past. However, MIB information cannot provide the full picture of an individual's health, especially if it does not include recent changes to their health status. Being able to access prescription and medical claims data in real time would be more informative and efficient—what can be accomplished using Milliman's IntelliScript® product. With an applicant's authorization, an insurance underwriter can use IntelliScript® to pull this data instantly to gain insights on an individual's past and current health status.

Countries that have universal health insurance may be able to develop a similar product that could be administered by the government or by a private insurance company administrating claims for the government. The ability to pull health records from a centralized social health insurance database would greatly assist the underwriting process.

In addition to using health claims and drug prescription histories for underwriting, this data can be used on an in-force block of business to track the block as it ages. These insights can help with new product and underwriting designs.

Vehicle driving history

Through the MIB, underwriters in the US can also access vehicle driving histories reported on past life insurance applications. This information is useful as there is correlation between risky driving habits and automobile crashes, which leads to higher mortality rates among drivers with frequent traffic violations. Updated driving histories can be pulled from a motor vehicle report (MVR) and be used to assess risk based on individuals' driving habits. In addition, risky driving behaviour may also be correlated with other behaviours that may lead to higher rates of mortality among risk drivers.

In the US and Canada, this information is made available to the public for a fee. One can obtain MVR for an individual by contacting the Department of Motor Vehicles (DMV) in each state. However, there are also third-party data vendors that aggregate this information across states to make it easier to obtain it—such as LexisNexis.[62]

As for the countries outside of the US and Canada, given our knowledge, it is not possible to gain access to a vehicle driving history unless your company is also selling car insurance and the client is already in your portfolio.

[62] LexisNexis - Motor Vehicle Records. Available at https://risk.lexisnexis.com/products/motor-vehicle-records

Clickstream and buying history

Starting from the definition ‘a clickstream is a record that contains data about a website user's clicks on a computer display screen via a mouse or touchpad. This type of information provides a visual trail of user activity with detailed feedback. Such data and related analysis facilitate market research and other scenarios concerning real-time user activity.’[63]

Can we go back to Bob for a moment? After coming home, he did a little research on the internet about life insurance. He entered **‘life insurance’** in the search engine and **clicked on the first result**. He was very delighted when the chat box opened, and a bot called Ana welcomed him, congratulated him for being a father, and offered some products that might suit his needs. He was surprised at first that she knew he was a father, but then realized **a few minutes ago he ordered diapers online**. After their highly personalized discussion, Bob said that **he needs time to think about the offer**. In the next few days, Bob saw advertisements for life insurance from this company while he was online, the same way you might see advertisements for a new TV while visiting Amazon.com. When **he clicked one of them** eventually, he did not have to explain all his needs from the beginning—the whole conversation and his needs were assigned to his IP address. Using accelerated underwriting, after a few other formalities, Bob signed his new policy.

How can we use it?

It is hard to imagine using clickstream in risk assessment or fraud detection; however, sales and marketing departments would happily use models that could recommend what kind of products or even which packages of products they should offer. Interest in such solutions grows especially in times of pandemic when the shift to digital sales is prominent. As the input data for the models is rather a set of our decisions and actions, the risk of discrimination is incredibly low, and probably that is why retail companies and portals like Netflix and Amazon are using them without regulatory backlash. We note, however, that many consumers explicitly forbid the tracking of their internet flow in their settings, and that needs to be respected.

Where can I get the data?

To start tracking clickstreams of your clients, you can use free vendors like Google Analytics. For more sophisticated analyses, you may need to contact paid vendors like Mixpanel, Kissmetrics, Amplitute, Heap, Google 360, or Adobe Cloud Marketing.[64]

Credit scoring / Purchasing power

Credit reports were developed for creditors and lenders to help them make decisions on whether to provide a line of credit to an individual. Credit reports contain more information than an individual’s single credit score. The report contains a summary of most current and past credit accounts associated with an individual and includes payment histories for each account. In the US, this information is gathered by three major credit bureaus: Equifax, TransUnion, and Experian.

While this information is traditionally used by creditors and lenders, it has many other uses. For example, in the US, credit checks are commonly used in employment and apartment/housing rental application screening processes. Credit reports are also used by US insurance companies to underwrite life, home, and auto insurance as credit scores correlate with many risks. Specifically, an individual’s credit score has been shown to correlate with mortality, which is why life insurance companies use it. There are many solutions currently on the market that use credit attributes in predictive models to predict mortality. These solutions are developed by credit bureaus and non-credit bureau companies such as Milliman IntelliScript and LexisNexis. The models these companies develop not only predict mortality, they also predict the risk that an individual will lapse a policy.

[63] What is a Clickstream? - Definition from Techopedia. Available at https://www.techopedia.com/definition/15403/clickstream

[64] A guide to clickstream data warehousing (stacktome.com). Available at https://stacktome.com/blog/a-guide-to-data-warehousing-clickstream-data#Vendors_-_Free

While credit reports offer many different purposes in the US, their use is highly regulated. In 1970, the Fair Credit Reporting Act (FCRA) was passed as a federal law to help protect individual consumers. It provides regulations on how data can be collected, maintained, used, and shared by credit bureaus. The act also entitles individuals to obtain a copy of their credit report from all three bureaus annually for no charge. This gives individuals the chance to review and dispute any discrepancies they see in their credit reports.

Credit bureaus also collect other information on individuals that are in public records. This information includes the following:

- Total amount of past due balances
- Number of accounts 90 days or more in last 24 months, plus derogatory public records (bankruptcies, liens, and judgments)
- Number of felony convictions
- Number of criminal convictions
- Professional license indicator

Life insurers also use this public record information when underwriting an individual for a new policy. Specifically, companies will assign individuals with a criminal record history as higher risk, which can result in higher premiums. If an individual has committed a severe crime, the insurers may refuse to issue a policy claiming the individual is too high of a risk to insure.

Telephone number / email address

Internal data which is easily accessed by insurance companies are telephone numbers and email addresses of potential clients. Usually, this is used only for the purpose of contact. Experiences coming from P&C business that ignore that information in the risk assessment can lead to loss making by not recognizing possible fraudulent claims.[65]

Fraud detection models can help to identify people with a higher probability of committing a fraud. The questions which we can ask ourselves are:

- If a person plans to make something illegal, is he/she willing to use his subscribed mobile number, or rather buy a prepaid sim-card?
- Would he/she use a main email account, or rather a newly created fake one?
- If the person calls us to report a claim (for example in a Long-Term Care product) from a telephone registered in another continent, should it make us suspicious?

Such indicators often help claim handling departments in tracing frauds, but thanks to models, claim handlers can obtain automated recommendations and focus on more sophisticated parts of their work, which cannot be handled by machines.

Moreover, while insurance companies collect contact information on their clients, they may not have the most current information. For example, if a person moves or changes phone numbers, it is likely they will not tell their insurance company right away. However, credit bureaus and other third-party data vendors collect contact information from a variety of different sources, which tend to have the most current contact information for an individual. An insurance company can purchase this information to keep their clients' contact information current.

Interactions with other companies

If insurance companies cooperate with a network of private hospitals, they may gain access to that network's client data. The insurance companies try to work with other industries, not only to find new sales channels, but to gain access to more data. Other examples could be the acquisition of banks, telecoms, third-party providers, or retail chains. Again, such sharing of data needs to be approved by all interested parties to avoid any data protection issues.

[65] Data in Paradise 2019 Session Recordings (carpe.io). Available at https://learn.carpe.io/dp2019

4.1.2 MARKET DATA

Financial data

1. Since life insurance is used as both a financial safety net for families and a product to build savings, market data can impact a policyholder's behaviour especially with regards to policy lapse and switching between funds. Possible external variables are listed in the table in section 3.3.1.
2. To quantify macro-economic variables so they can be used in a model, we must use indicators. A little creativity can be used to produce these indicators, but some examples are listed below:
   - Indicators like the consumer confidence index (CCI) or results from a consumer demand survey can be used to gauge buyer confidence.
   - CCI equates to the level of optimism with regards to the current economy consumers are expressing through their daily spending and saving activities. For reference, in the US, CCI surveys are done by The Conference Board, while in Japan the survey is done by the Cabinet Office.
   - Consumer demand surveys gauge demand for certain goods like cars, TVs, PCs, furniture in a specified time (in three months, six months, 12 months). This can measure purchasing behaviour in the near future or the optimism to buy in the future.
   - Indicators like the consumer price index (CPI) or personal consumption expenditures (PCE) can be used to gauge inflation.
   - CPI and PCE are similar but differ in the formula, data, and weights they use. For reference, in Japan, the CPI is calculated by the Ministry of Internal Affairs, and in the US, the CPI is calculated by the Bureau of Labour Statistics.
   - Indicators like the House Price Index (HPI) can be used to quantify housing price development.
   - If interested in general indicators of the economy, market index funds like the S&P 500 can be used.
3. Economic variables should be chosen depending on each individual use case and what economic data is available. Economic variables like gross domestic product (GDP) can be accessed in the World Bank database using an API26, while stock data can be accessed on Yahoo Finance via API27.

Web aggregators and competitors' prices

Web aggregators are relatively new in the life insurance business and, in some countries, it is still impossible to get a quote and buy a policy online. However, some markets where this is possible include The Netherlands, UK, and US. Thanks to that, it is possible to get an idea of competitors' prices and to use them in beneficial ways:

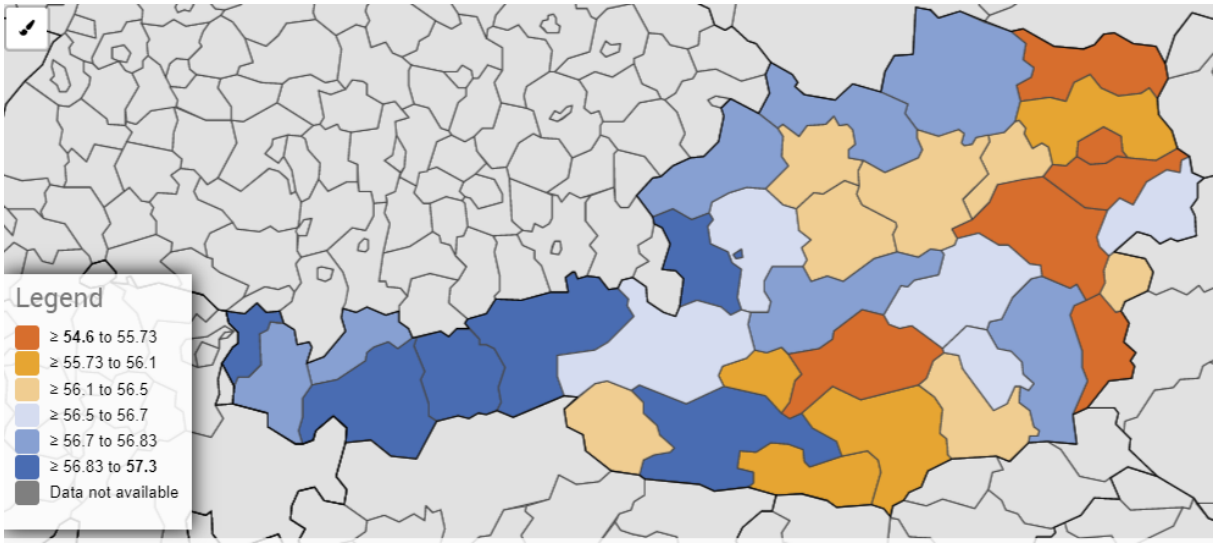
- When your own products are in a web aggregator, you not only add a new sales channel, but also gain access to conversion rate statistics which show what percentages of clients bought a policy after quotation. Thanks to this information, it is possible to create price elasticity models.
- There is no doubt that price elasticity changes along with prices on the market. Knowing what prices other companies are offering helps to improve models and to react in real time to market changes.
- Lastly, when an insurance company launches a new product, very often it is not easy to set the right price due to the lack of historical data. Using information from web aggregators can fill these knowledge gaps.

We should mention that price elasticity models are currently under scrutiny by regulators in the UK because of a potential lack of fairness. Please check current regulations in your country before applying such a solution.

4.1.3 GEOGRAPHIC INFORMATION SYSTEM

Using GIS in life pricing models

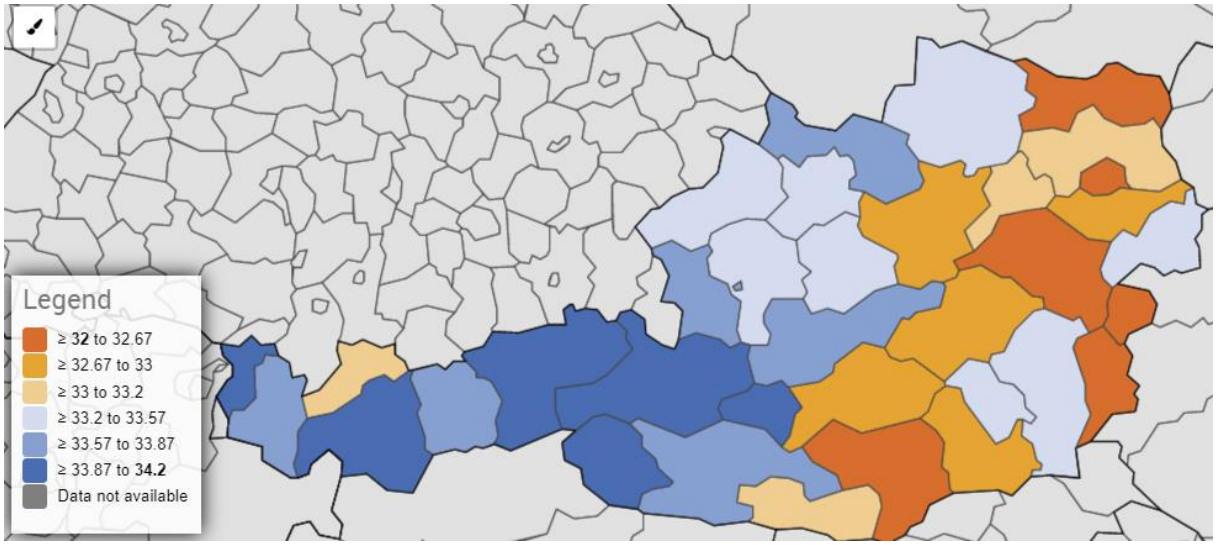
FIGURE 12: GIS PRICING EXAMPLE – LIFE EXPECTANCY OF AUSTRIAN FEMALES, AGED 30, IN 2030



Life expectancy for females who will reach 30 years in year 2030, statistics published by Eurostat (European Statistical Office). Results split by NUTS 3 regions in Austria. Screenshot from: Statistics | Eurostat (europa.eu)

On the map above we can see that life expectancy in the western part of Austria (close to Switzerland and the Alps) is almost three years longer than in the eastern part, including Vienna, the capital (Austria is used here as an example; similar scenarios can be found in other countries). These results are not caused by different age or sex groups in the region, because we show life expectancy for both in the map. Nor is it random noise in the number of deaths because we are looking at life expectancy, not death rate. A similar picture can be seen for males aged 50 in 2030.

FIGURE 13: GIS PRICING EXAMPLE – LIFE EXPECTANCY OF AUSTRIAN MALES, AGED 50, IN 2030



Questions we can ask ourselves are:

- What is the reason that people live longer in the western part of the country?
- Is this reason already explained by other variables to which we have access like purchasing power or lab results in the age of underwriting?
- If not, are we blind to these reasons?
- If the second option is true, do we really want to offer the same price in the areas we expect people will die later? Or, do we really want to distribute marketing resources evenly across the country even though we see a less risky portfolio in some specific area?
- **And most importantly, if we are not sure which possibility is true, can we afford not to test these hypotheses?**

The usage of a residence address in pricing lines of business such as property or motor insurance does not surprise anyone. It is obvious that the risk of floods or the density of cars is higher in some areas. As a result, prices vary significantly. For example, people in big cities pay almost two times more for their car insurance (it happens even though the place where the car is registered is not necessarily in the area the car will be used, however, increased usage of IoT and telematics help to assess real locations). The P&C industry faces problems with geographical modelling as well, but they have found good solutions in recent years. Our conclusions, based on these experiences, are:

- It is possible to use geographical variables in models using credibility theory or machine learning methods.
- Instead of using multi-level categorical regional variables (like postal codes), it is possible to use external variables correlated with risk.
- The risk of a car collision ending in a bodily injury in a big city increases about two times compared to villages (this risk is more significant for younger people who are at a higher risk of dying in an accident than dying of illness). Thus, it is undeniable that mortality risk is different in different areas because it includes the risk of dying from car collisions and other regional factors.

What is the problem with geography?

The fact that life insurance companies rarely use geographical variables in their pricing models is not without reason. In Poland, for example, there are 16 country regions, 380 counties, and almost 20,000 postal codes. If we use country regions (voivodships) in our models, the data may not give us enough insight, as it does not differentiate sufficiently between higher-populated areas and less-populated ones. On the other hand, if you would like to have another price in each postal code it will not be possible (and reasonable), especially when claims frequency in the life business is much lower than in the P&C industry and thus do not produce enough data to train the right risk coefficient for each postal code..

We must also remember that during a 20-year contract a lot of things can change in the life of the policyholder and home address is not an exception here. Moving policyholders make it difficult to attach risks to a certain region. However, we can look for similarities in motor pricing models where different home addresses and car usage areas do not cause such a big problem (keep in mind that the duration of car insurance is much shorter). Practical solutions for this issue were discussed in section 4.1.1.4. For variables like pollution or climate change, we can try to estimate the level of external variables we expect to see in the next few years.

Another argument against modelling geography is model simplicity. Pricing life insurance products is relatively simple and both actuaries and regulators like it this way. Adding new variables can create uncomfortable questions.

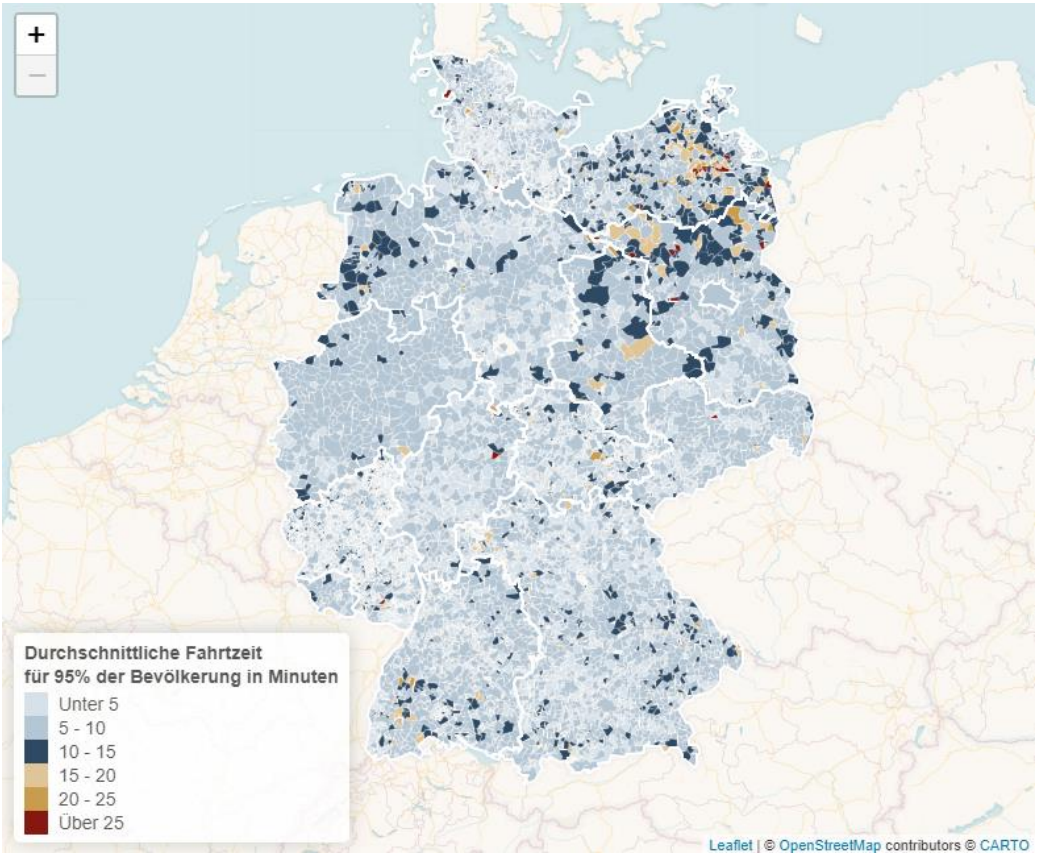
If that is not enough, a lot of doubts come up when we consider regulatory and ethical issues, because by using geography we could inadvertently cause discriminatory biases. We must consider these doubts and others.

However, we believe it may be worth asking these questions. Thanks to more sophisticated models and better risk stratification, it is possible to find better risk sub-portfolios and, as a result, significantly improve profitability and/or increase the volume of sales through more competitive pricing.

What are the reasons for different mortalities across countries?

There are many reasons mortality rates differ across countries. The fact that a higher frequency of car accidents with bodily injuries occurs in big cities is one example. The map below gives another example. The average time an ambulance reaches the home of a patient varies between regions in Germany, somewhere between 5 and 28 minutes.[66]

FIGURE 14: GIS EXAMPLE - AVERAGE WAIT TIMES FOR AMBULANCE SERVICES IN GERMANY



Is there any chance the mortality will be higher in regions where you need to wait longer for an ambulance? What about the availability of a telecom signal when you need to call an ambulance? Again, with this example it seems reasonable to test whether the areas without such signals have higher mortality rates. External variables can offer a snapshot of reality as it stands now, which can change in the future. For this reason, it may be safer to use such variables only in pricing short-term contracts.

And what about the different treatments of disability status in different regions of the country? Death is death, but which health status should be classified as a permanent disability? These definitions can vary between regions due to non-centralized court and administration systems (Switzerland, for example) and, for this reason, the probability of paying a claim due to permanent disability of the insured varies between regions.

Other variables can be correlated with mortality, such as:

- Weather (humidity, number of rainy days, number of snowy days, average temperatures)
- Levels of pollution
- Health care shortages, number and distribution of hospitals
- Smoking, drinking habits
- Education levels

[66] Rettungsdienste brauchen im Osten länger - Institut der deutschen Wirtschaft (iwkoeln.de). Available at https://www.iwkoeln.de/presse/iw-nachrichten/henry-goecke-rettungsdienste-brauchen-im-osten-laenger.html

- Criminality (is the risk of dying higher if we live in a dangerous neighbourhood?)
- Purchasing power
- Proximity to healthy restaurants
- Proximity to gyms

The analysis of the correlation between some chosen socioeconomic statistics and mortality was performed in the paper 'Modelling and Forecasting Cause-of-Death Mortality by Socioeconomic Factors (soa.org)' created by Milliman specialists in Paris and the US.[67]

Where can I get the data?

You can start with the national statistics of your country. If you are in Europe, Eurostat can be a good option. Statistics available on the NUTS 3 level (Nomenclature of Territorial Units for Statistics) are gathered for each European country. The most granular data to be found in the US is the National Census where you can find detailed statistics even on the level of Census Blocks, essentially, what equates to the size of a few houses. If you want to keep up with freely available data sources, you can use Open Street Maps to create API connections with a number of points of interest in the neighbourhood (for example the number of hospitals in a given postal code) or the proximity to some points of interest (like the distance to the next hospital from your home address). If the above sounds a little bit overwhelming, you can always get in touch with paid providers.

Using GIS in other life models

Here the situation is easier. Models such as lapse analysis, fraud detection, or a client's segmentation are less regulated because they are less likely to lead to any adverse action. For example, if we see that the probability of lapsing is higher in some areas, we may simply get in touch with policyholders living there and offer them discounts or cross-sales of other products (their life stage might have changed).

It is impossible to list all possible variables here. The most important task is to start testing.

4.2 Regulation and ethics for external data

Before we use any machine learning models, we need to make sure that all interested parties are comfortable with the solution including data owners, clients, regulators, and the insurance company itself. AI now surrounds, starting from Netflix and Amazon recommendations, through ordering a taxi with an app, and when Siri recognizes our voice. When used properly, AI can enrich humans' lives, for instance by improving health care, protecting the environment, or enhancing how humans interact with each other and the world around them.

Regulators struggle to keep up with the AI trend and the speed of its development. That is why the insurance industry must not only obey all the rules but be a leader of the data protection and ethical usage of AI. At the end of the day, behaving in line with professional standards like the codes of conduct in the US or in the EU, and fulfilling responsibilities to society, clients, and employers, is the fundamental duty of each actuary.

'The Code of Professional Conduct sets forth what it means for an actuary to act as a professional. It identifies the responsibilities that actuaries have to the public, to their clients and employers, and to the actuarial profession.'[68]

It is out of the scope of this paper to describe fully this extraordinarily complex matter. For most detailed information about regulation and ethics for AI, we encourage you to reach for another Milliman paper called 'Artificial Intelligence: The ethical use of AI in the life insurance sector'.[69] However, we will summarize here the most important implications from the perspective of external data, based largely on the splendid work done by Milliman specialists from the UK.

[67] Modeling and Forecasting Cause-of-Death Mortality by Socioeconomic Factors (soa.org). Available at https://www.soa.org/globalassets/assets/files/resources/research-report/2021/modeling-and-forecasting-cause-of-death-mortality-by-socioeconomic-factors.pdf

[68] Code of Professional Conduct | SOA. Available at https://www.soa.org/about/governance/about-code-of-professional-conduct

[69] Artificial Intelligence: The ethical use of AI in the life insurance sector (milliman.com). Available at https://uk.milliman.com/-/media/milliman/pdfs/2020-articles/articles/11-24-20_ethics-ai_20201117.ashx

4.2.1 WHAT ARE THE KEY CONCERNS?

- Machine learning models aim to discriminate—that is how they have predictive power. The features that such models use to determine predictions are raising ethical questions around bias, fairness, and discrimination against groups and individuals.
- AI systems are increasingly allowing value to be extracted from unstructured data sources that were previously inaccessible, including some sources that individuals might not to choose to share with insurers (e.g., social media data). This raises ethical questions around data privacy and data security.
- Machine learning models can be extremely complex (e.g., neural networks) and so it is hard to understand how they are making decisions (i.e., they are black boxes). When new data sources are added, it may be difficult to assess which variables are being used in decisions. There is potential for seemingly innocuous data to act as a proxy for variables which insurers are not allowed to use. For example, a model could offer higher premiums to individuals who work in construction and who watch war movies, perhaps as a proxy for being male.
- Additional data could allow for hyper-personalized pricing specific to an individual's circumstances, eroding the risk-sharing element of insurance. At the extreme, this could lead to having uninsurable sections of society.
- Furthermore, data could also be used to assess price elasticity and identify individuals who will accept higher prices, potentially targeting vulnerable individuals who are less savvy at shopping around for a better deal. Such sub-portfolios may include both those who are earning more money and older customers who are not so experienced in comparing their quotes on the internet.

Moreover, vulnerable and less privileged groups such as people without access to smartphones or people, and who are less aware that publishing something online could lead to changing a price of their premium (as described in section about social media) can be at risk of unfair treatment.

4.2.2 CURRENT REGULATIONS

AI models are covered by many different types of existing regulations. When building an AI model, we should consider the following:

- Equality legislation, such as the EU Gender Directive which prohibits pricing based on gender
- Data Protection legislation, such as the EU GDPR which limits which data can be used and how
- Insurance industry-specific regulation, such as Financial Conduct Authority (FCA) rules on fair pricing practices
- Any AI specific legislation, e.g., the European Commission's proposed Regulation Laying Down Harmonized Rules on Artificial Intelligence

There can be additional guidance on creating ethical AI models which may not be binding but may be viewed as best practices. Insurers may look to the following types of organizations:

- Financial supervisory bodies – voluntary guidance or expectations on firms may exist before specific regulation is in place
- Data regulation authorities – e.g., the Information Commissioner's Office in the UK
- Professional bodies – e.g., the UK Institute & Faculty of Actuaries has a guide for members creating AI models
- International bodies – e.g., the European Commission has published extensive guidelines on ethics that are of relevance to U.K. and European insurers
- Government advisory bodies – e.g., the UK has a Centre for Data Ethics and Innovation
- Relevant independent research bodies – e.g., the Alan Turing Institute and the Ada Lovelace Institute in the UK

Firms should investigate what guidance is available which might apply to their use case.

The concept of discrimination is set out in human rights such as the EU's Charter of Fundamental Rights, which broadly prohibits discrimination on the grounds of human characteristics 'such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation.'[70] It is important to mention that the discussion about discrimination and fairness in insurance is nothing new. Documents such as the Equality Act and European Gender Directive are about 10 years old.

In the U.K., insurers can differentiate on certain explainable attributes where their reasoning is based on statistical evidence but cannot do so when it is unjustified or unexplainable. For example, discrimination via differential treatments of the grounds of age, is not discrimination under the Equality Act 2010, provided it is shown to be 'a proportionate means of achieving a legitimate aim.' Thus, price can vary by age, or lives can be bounded into groups based on age, where this genuinely reflects risks or costs. Indirect discrimination is not discriminatory provided it can be justified as a proportionate means of achieving a legitimate aim. On the other hand, European Gender Directive states clearly that from 2012 onwards, EU insurers cannot use gender as a rating factor in their models.

The insurance companies often do not even gather sensitive information such as that pertaining to race and, as a result, it is sometimes hard to prove their models are not indirectly discriminative (for example by using proxy variables highly correlated with discriminatory variables). This presents a challenge for insurers; that means insurers need to differentiate—they cannot set the same premium for all lives. Instead, insurers must determine that they are comfortable with how the model is differentiating (and to document this), rather than simply claiming it is differentiating.

4.2.3 SPECIFIC REGULATIONS ABOUT EXTERNAL DATA

As the subject of external data use in insurance modelling is most developed in the US, due to the fact that the US is where the largest data is available, the most specified regulations have been created. New York Circular Letter issued in 2019[71] [72] on the one hand encourages usage of modern technologies, but also expresses concerns that external data can lead to discrimination and unfair practices. Therefore, the letter gives the guidance that external data can be used only if the insurance company can provide evidence documenting that there are no indications that a model is discriminatory. As the letter was followed by the Colorado Senate Bill in 2021,[73] which gives a similar message, it can be expected that the following regulations may appear in other parts of the world. As for the EU, the main limitations for the use of external data results from the GDPR.

4.2.4 POSSIBLE SOLUTIONS

It is critical for modellers to understand where bias and discrimination could arise in their model, and that they should be aware that there may be no perfect solution, with trade-offs between accuracy (complexity) and bias in modelling.

The good news is that there are solutions that can investigate a range of fairness measures in machine learning models including Aequitas (Aequitas, 2020), TensorBoard's What-If-Tool (TensorFlow, 2020), and IBM's AI Fairness 360 (IBM, 2020). For example, the latter is an open-source tool kit which can access more than 70 fairness metrics and can be used in common programming languages such as R and Python.

[70] Article 21 - Non-discrimination | European Union Agency for Fundamental Rights (europa.eu). Available at https://fra.europa.eu/en/eu-charter/article/21-non-discrimination

[71] SC-Publication-NYDFS-Addresses-Use-of-External-Consumer-Data-in-Life-Insurance-Underwriting.pdf (sullcrom.com). Available at https://www.sullcrom.com/files/upload/SC-Publication-NYDFS-Addresses-Use-of-External-Consumer-Data-in-Life-Insurance-Underwriting.pdf#:~:text=On%20January%2018%2C%202019%2C%20the%20New%20York%20State,and%20information%20sources%20in%20underwriting%20for%20life%20insurance.1

[72] NY DFS Delivers an Icy Blast to Insurers Using External Data Sources and Algorithmic Underwriting | Carlton Fields. Available at https://www.carltonfields.com/insights/publications/2019/ny-dfs-delivers-an-icy-blast-to-insurers-using-ext

[73] Colorado Law Bars Insurance Discrimination By Data - InsuranceNewsNet. Available at https://insurancenewsnet.com/innarticle/colorado-law-bars-insurance-discrimination-by-data

To answer the issue of not having discriminatory variables (like race), which makes assessment of fairness hard, AI again can help by imputing race of a person based on features like the name of a person.[74] Using external data about a person's race can help as well.[75] At the same time, such models must be used with caution to avoid their biases!

Furthermore, you can find a case study presenting an analysis that tested a model to confirm that it showed no indication of it being racially discriminatory in the paper called 'Testing Milliman Advanced Risk Adjuster models for racial bias.'[76]

Apart from the need to prove fairness of their models and used input data, insurance companies must care for topics such as data security to ensure that externally sourced data comes with consent of the customer and explainability of the models.

4.2.5 MORE RESOURCES

For more guidance, we strongly encourage you to reach for a much more detailed paper created by Milliman specialists mentioned before69 and for a summary of the report by EIOPA's Consultative Expert Group on Digital Ethics in Insurance, created by the same Milliman UK team, called 'Artificial Intelligence governance principles.'[77] If you want to further explore the subject of regulations, you may want to look at the 'General Data Protection Regulation'[78] (GDPR) in the EU, 'Health Insurance Portability and Accountability Act of 1996'[79] (HIPPA) in the US, and the proposed 'Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence' (Artificial Intelligence Act).[80] You can also refer to the UNCTAD website[81] to find out which exact Data Protection and Data Privacy Legislation is in force for each country across the globe.

[74] When Race/Ethnicity Data Are Lacking: Using Advanced Indirect Estimation Methods to Measure Disparities | RAND. Available at https://www.rand.org/pubs/research_reports/RR1162.html

[75] Using publicly available information to proxy for unidentified race and ethnicity A methodology and assessment 201409_cfpb_report_proxy-methodology.pdf (consumerfinance.gov). Available at https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

[76] Testing Milliman Advanced Risk Adjuster models for racial bias. Available at https://www.milliman.com/-/media/milliman/pdfs/2020-articles/articles/9-1-20-testing-mara-models-racial-bias.ashx

[77] Artificial intelligence governance principles (milliman.com). Available at https://www.milliman.com/-/media/milliman/pdfs/2021-articles/7-30-21-artificial-intelligence-governance-principles.ashx

[78] General Data Protection Regulation (GDPR) – Official Legal Text (gdpr-info.eu). Available at https://gdpr-info.eu/

[79] Health Insurance Portability and Accountability Act of 1996 (HIPAA) | CDC. Available at https://www.cdc.gov/phlp/publications/topic/hipaa.html

[80] Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) resource.html (europa.eu). Available at https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

[81] Data Protection and Privacy Legislation Worldwide | UNCTAD. Available at https://unctad.org/page/data-protection-and-privacy-legislation-worldwide

How Can Milliman Help?

Keep in mind that Milliman has teams specialized in wearables usage, gathering and analysing health and credit scoring data, gathering GIS statistics, creating various Data Science solutions, and analysing regulatory and ethical aspects of using external data and Data Science. Milliman can help you with all aspects of your Big Data projects and Data Science needs, including advice on:

- Recommending best practice frameworks for Data Science processes
- Collecting and processing data (both internal and external)
- Showing suitable tools and techniques for circumstances
- Implementing Data Science solutions
- Assisting in model development, validation, testing, and reporting
- Understanding the implications of results
- Advising on constraints and practical challenges

For further information, please contact your usual Milliman consultant or those listed below. We would appreciate your feedback related to this paper or about what other topics you would like us to cover in the future.

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT
Bartosz Gaweda
bartosz.gaweda@milliman.com

Christoph Krischanitz
christoph.krischanitz@milliman.com

Remi Bellina
remi.bellina@milliman.com

Jeff Anderson
jeff.anderson@milliman.com

Joe Long
joe.long@milliman.com

Noriyuki Kogo
noriyuki.kogo@milliman.com

Saiki Justin Makino
saiki.makino@milliman.com

Scott Chow
scott.chow@milliman.com

© 2022 Milliman, Inc. All Rights Reserved. The materials in this document represent the opinion of the authors and are not representative of the views of Milliman, Inc. Milliman does not certify the information, nor does it guarantee the accuracy and completeness of such information. Use of such information is voluntary and should not be relied upon unless an independent review of its accuracy and completeness has been performed. Materials may not be reproduced without the express consent of Milliman.