Big data and risk classification

Understanding the actuarial and social issues

July 2022

Canadian Institute of Actuaries
Institut canadien des actuaires

Executive summary

Risk classification is an important function within the insurance rate-making process. It ensures that insurance policy owners are charged a price for their coverage that is reflective of the underlying risk being insured. The recent emergence of big data has allowed insurers to make increasingly more refined pricing decisions.

However, concerns have been raised that the use of big data to generate a reflection of the underlying risk could lead to social inequities. As such, there have been calls to limit or restrict the use of big data in risk classification decisions.

The Canadian Institute of Actuaries (CIA) believes that the use of big data to gain insight serves an important place in the healthy functioning of insurance markets. Accordingly, its use should be permitted, subject to critically important limitations relating to protected classes under human rights legislation, such as race, sexual identity, and religious expression.

Further, we believe that such data creates new opportunities to understand risk and its underlying contributing factors through scientific study. With improved insights regarding risk and its contributing factors, we can collectively make progress that reduces risk and lowers insurance costs.

Big data: Large, complex datasets of aggregated information that are gathered rapidly or on an ongoing basis, which are analyzed and anonymized to identify broad trends and patterns. Actuarial work involves examining data on an aggregate basis to identify and quantify trends, and is not concerned with the identity of the contributors to that data.

Fairness: The word "fair" carries with it a number of different connotations. "Actuarial fairness" means that policy owners are charged a price for coverage that is reflective of the underlying risk being insured. It is not meant to imply "fairness" in the sense of affordability, or other considerations, which can be referred to as "social fairness."

Classifying risks

Life is full of risk. Whether it's injury or loss of life from sickness, accident, or disease, or property damage from natural or human disasters, most members of society are vulnerable to the financial consequences of unexpected adverse events. People have built mechanisms to help mitigate, at least in part, the unfavourable financial or personal impacts of such adverse events should they occur. Insurance is one such mechanism.

Insurers place individual risks into groups with similar characteristics and probabilities called "risk classes." While it is not possible for any insurer to predict which individual risk in the risk class will incur a claim, it is possible to predict, with varying degrees of precision, the claim costs for a risk class as a whole.

To promote equity and actuarial fairness within the insurance system, insurers charge different prices according to risk classes. The higher the anticipated claim costs per individual risk within a risk class, the higher the price and vice versa.

Insurers have always relied on the extensive use of data in designing and maintaining insurance systems and risk classification systems. This helps to ensure those systems are sustainable and charge an amount for coverage that is appropriately reflective of the underlying risk. Data also makes insurance more broadly available, in terms of providing coverages for risks that otherwise would be difficult to price.

In recent years, there is ever-increasing availability of big data made possible by our digital lives. Important issues have arisen about the use of such data for the purposes of risk classification. Those issues include the intended use of data, data privacy, correlation to risk, unfair or unknown bias within data, and a host of other practical, ethical, and legal issues.

Big data can be viewed as an extension of insurers making use of data with additional potential issues and benefits. Restrictions on the use of certain risk characteristics could adversely impact the availability or price of insurance. As actuaries, we believe in thoughtful innovation and evolution in the use of big data, while ensuring that the public interest is at the forefront of insurance and policymaking.

Risk pooling

Insurance is a way to transfer financial risk. For example, a homeowner could lose their house due to a fire, a storm, or some other peril. Even if the probability of such an event occurring is small, it is not a risk that most people would choose to take.

Here's an example of how this financial risk transfer takes place. Imagine that we are dealing with the risk of a house being completely destroyed by fire. Without insurance, the homeowner's outcomes are either:

a) incur no catastrophic loss if no fire happens,

or

b) suffer a catastrophic loss if a fire happens.

The homeowner faces a lot of risk in the case of outcome b. With enough insurance, this is reduced to one outcome – a manageable cost (which is the insurance premium) with no risk of catastrophic loss. By purchasing an insurance policy, the owner of the house mitigates their risk by exchanging one set of outcomes (an uncertain but possibly catastrophic loss) for another (a certain but manageable cost). The financial risk is effectively transferred to the insurance company.

The insurance company manages the risk by taking advantage of the "law of large numbers." Events such as a house fire are considered to be random, meaning that there is no reliable way to predict when they will take place or who will be affected. Through the collection of claim events that occur over time, it is possible to estimate how frequently these events occur on average.

For the insurance company, the claim cost is still random, but by insuring a large number of houses, the insurance company can predict with greater certainty that the total amount of claims they will pay will be sufficiently close to the historical average to make the insurance arrangement viable.

This is referred to as "risk pooling" – a mechanism where individual risks are collected together so that the claims from the unfortunate few can be covered by the premiums collected from everyone in the pool.

Risk classification

Example 1 of risk pooling for houses was designed to be very simple. In reality, there can be significant differences between the houses insured – such as size and building features – that will affect the cost associated with an event such as a fire. It would obviously not be appropriate to charge the same cost on all insurance policies in this case.

Risk classification is used when the price charged for insurance reflects underlying characteristics of the individual or object being insured. Some examples include:

- Females have higher life expectancy than males; they are charged lower rates for life insurance.
- Smokers have higher mortality than non-smokers; they are charged higher rates for life insurance.
- Young drivers have higher accident rates than older drivers; they are charged higher rates for auto insurance.
- Individuals in certain occupations (such as logging) have a higher risk of injury than others (such as accountants); they are charged higher rates for disability insurance.

Risk classification allows insurance companies to charge a price for the financial risk transfer that better reflects the nature of the underlying risk.

If risk classification did not take place, there would be significant cross-subsidization when it comes to the pricing of insurance. Cross-subsidization means that one class of individuals pays more to transfer their financial risk than the true cost and another class pays less (see Example 2).

Example 1: Risk pooling for houses

Suppose that houses are worth $500,000 on average, and the probability of one being completely destroyed due to a fire is 0.1%.

An insurance company could then charge a cost of 0.1% x $500,000 = $500 to each customer to cover the claim. If they insure 100,000 houses, they will collect $500 x 100,000 customers = $50 million.

On average, they will expect to see 0.1% x 100,000 = 100 houses lost to fire, which costs them $500,000 x 100 = $50 million, which matches the premiums collected.

Of course, there is no guarantee that the 0.1% probability will be exactly realized. However, by insuring a large number of houses, the insurance company increases the likelihood that the outcome will be close to the estimate. Further, insurers will need to collect adequate premiums to cover other costs related to selling and servicing policies, claim settlement processes, profit, etc.

Example 2: Risk classification for houses

Imagine that 20% of houses have wood stoves and 80% have gas furnaces. Suppose also that the probability of having a claim due to a fire is 0.3% for houses with wood stoves and 0.05% for houses with gas furnaces.

It can be shown that the probability across all houses is still 0.1%, so the same $500 premium per house would still cover the expected claim costs. However, the houses with gas furnaces would pay $40 million in aggregate to cover $20 million in claim costs, while the houses with wood stoves would pay $10 million in premiums to cover $30 million in claim costs.

In this example, the homeowners with gas furnaces would be subsidizing those with wood stoves.

We also see this effect in life insurance, which pays a benefit upon the death of the insured individual. It is well known that the probability of dying increases as individuals get older. As a result, it would not be actuarially fair to charge the same insurance cost to everyone regardless of age, since the risk of death increases as the insured individual gets older.

If life insurance rates did not vary by age, then younger individuals would pay significantly more than the true cost, and older individuals would pay less. The result would be that younger individuals would elect not to purchase insurance, the premiums collected on older individuals would be insufficient, and the entire insurance model would no longer function properly.

Risk classification is the term used to describe the identification of risk characteristics that influence claim outcomes, and the determination of cost differences charged to individual insureds to maintain an adequate balance between total claim outcomes and premiums collected.

The identified characteristics used in determining these cost differences are called “rating factors.” Examples of rating factors for life insurance include age, sex, smoking status, lifestyle, occupation, and health status. For auto insurance, these include the type and age of vehicle, the driver’s number of years licensed, history of at-fault accidents, driving violations, and the location where the vehicle is garaged. In each case, these rating factors correspond to identified risks that have been shown to increase or decrease the possibility and/or amount of a claim.

Anti-selection

A functioning insurance arrangement depends on transparency of information between the insurance company and the individual being insured. This means that insurance applicants should disclose any facts they are aware of that could affect the price of the insurance. Otherwise, the applicant may have the opportunity to "select" against the company.

Imagine someone is applying for life insurance and has recently been diagnosed with a serious form of cancer. If the insurance company knows this information, they will charge a higher cost for the insurance to reflect the increased risk, or perhaps decline to issue the insurance at all because the cost would be unaffordable.

However, if the insurance company does not know, or does not receive disclosure of this information, then the applicant could obtain insurance at standard rates, even though they represent a significantly elevated risk. The insurance company would effectively be undercharging for the risk they are assuming. Anti-selection can also occur when risk classification is not sufficiently granular, as in Example 3.

It should be noted that the discussion of this and other issues is based on an open and voluntary market for insurance.

Example 3: Anti-selection for houses

Suppose that Company A charges different costs to houses with wood stoves and those with gas furnaces, while Company B charges the same cost to everyone.

Homeowners with wood stoves will more likely purchase their insurance from Company B because it is less expensive for them. Company B could then find that their proportion of wood stove business is no longer 20% but something significantly higher, in which case they would not have collected enough premiums to cover their expected claim costs.

In the extreme, this situation could lead to insurance companies becoming insolvent, which in turn reduces the opportunity for homeowners to transfer financial risks.

Big data and risk classification

Collecting personalized data

The term big data is often used in conjunction with the technology that allows for the collection of significant amounts of information on individuals and the ability to use that data to develop insights about the individual. Examples of big data include your Google search history and Facebook advertisements (see Example 4). The benefit of the use of big data for the user is a more personalized and relevant user experience.

Insurers have been using data for many decades to better assess risk of individuals and the overall risk pool; the benefit is the ability to offer more types of insurance (e.g., overland flood insurance or living benefits insurance) as well as more granular pricing. Insurers are now leveraging big data for the new opportunities it presents. Some examples include using health data from wearables (like Fitbits) and driving data from vehicle monitoring devices (such as telematics devices in the automobile or mobile phone apps).

For insurance, big data has the potential to significantly improve the classification of risk as it allows the insurer to use information that more accurately aligns with the true underlying risk. For the consumer, big data can ease the sharing of data to the insurer, allow for more personalized products and services from the insurer, and help the consumer better understand the costs of insurance. This may lead insurers to improve existing risk classification methods or to develop new classifications.

Example 4: Wearables and insurance

In using data from personal health devices (wearables) for insurance, a variety of data points are collected: steps, heart rate, and body temperature.

Data can track activities from moderate to high activity, such as a vigorous walk or run.

Similar to good driving habits, good physical fitness levels can be used to classify an individual as lower risk for early death or developing certain diseases, for example.

This data can be used to validate that a physical activity regime is being adhered to, which could be beneficial for higher risk groups, such as diabetics, to demonstrate they are maintaining or improving their health.

This validation of the individual's health activity regime can be used to place them into a lower risk category, giving the individual more control of their insurance pricing.

A good example of this new type of data used by auto insurers comes from vehicle driving data, which is collected from devices such as smartphones, that can monitor driving habits. While driving risk classification looks at historical driving incidents such as accidents or traffic violations, it may be overly punitive on new drivers who practice safe driving behaviours but who do not have enough history for those behaviours to be captured accurately. By using actual driving experience, new types of data such as frequency of hard braking, speeding, or excessive acceleration can be used to better reflect risky driving behaviours.

Using personalized data

As with all uses of big data, there is a necessary balance between the benefits to the user and the potential for misuse. The collection of additional data may result in a better assessment of the risk, but it may also compel the user to share information that is not relevant to that assessment.

The ability to forecast a risk with certainty will always be elusive, and while big data will improve forecasting accuracy, it alone will never completely eliminate forecasting error.

Actuarial work is primarily concerned with identifying trends and risk classification factors within the data, so even though the user may end up sharing information about themselves, the actuarial analysis does not tie this information to any specific individual. Once this analysis is completed, and the end product (e.g., a rating system) is created, then individual factors can be taken in account for the purpose of insuring that individual. Prior to that, aggregate data without individual identification is sufficient.

Bias in risk classification

Fundamentally, risk classification provides positive benefits to society by measuring the costs of risks – and therefore enabling financial risk transfer activities. This process also provides information to insurance consumers that helps them measure the "riskiness" inherent in particular activities. With this type of information, consumers can in certain circumstances modify their behaviour to reduce or avoid their exposures to these risks.

Usage-based insurance refers to risk classification programs that incorporate the use of technology to gather detailed information, such as regarding an insured vehicle's usage. These programs may collect details such as the starting point of a trip, the end point of the trip, the start time and duration of the trip, as well as a multitude of characteristics regarding the speed and acceleration rates measured at millisecond intervals along the way. Using this information, actuaries employ statistical methodologies to identify relationships between these particular trip traits and the accidents that occur during these trips.

This analysis provides information regarding the risks of driving behaviours, such as exceeding posted speed limits. The use of this information in a risk classification system would reward safer driving behaviours with lower insurance rates and discourage higher-risk driving behaviours with resultant higher insurance rates. From a social fairness perspective, most would also consider this outcome "fair" without controversy given that speeding is an easily avoided behaviour, and the social benefits of lower rates of speeding are intuitive and apparent. There is a long history of research studies and public education campaigns that have accompanied the introduction and reinforcement of speed limits on public roads and highways. Many of these have been accompanied by very graphic and visual representations of the dangers of speeding. Accordingly, there is a strong public awareness of the dangers of speeding, and so its use in the insurance risk classification program is likely to be uncontroversial.

The situation, however, is less clear in the case where a characteristic is both less controllable and where the interaction with driving risk is less intuitively clear. Take, for example, the start time of an insured driving trip. Studies on usage-based insurance data have shown that insured trips starting after midnight pose a higher risk of accident than a trip that begins at two in the afternoon. Following our speeding example, a natural insurance outcome might be to surcharge trips that begin between 12 a.m. and 4 a.m. and to provide an offsetting discount for trips that start in the period between 2 p.m. and 4 p.m. From a societal perspective, we

know that such a program would reduce trips in higher risk time periods to the extent they are avoidable, and so from an aggregate view, introducing this risk element would save lives and reduce accident costs.

It is helpful to understand the context for the higher-risk relationship. Does the higher rate of accident occur because drivers are tired? Is it because this time also coincides with the closure of facilities where alcohol has been served? Actuarial studies can address these questions with some certainty (e.g., perhaps by examining the intersection of accident rates after visits to bars in this time window), but the public discourse will benefit if the relationship can be explained. Due to other factors, however, such as privacy regulations or the fact that some interactions cannot ever be fully explained, we may not have a clear view.

Insurers do not require that causality be explained, partly because in many cases it is impossible to determine, although it is important that the insurer is convinced that the risk characteristic is correlated to the true (but unknown) risk drivers. In some cases, however, even where we have a clear causal relationship between a particular insured characteristic and risk-related outcomes, the use of the characteristic may impact a particularly vulnerable segment of society adversely.

It is possible that some individuals who must drive at late hours in the early morning are, for example, staff working overnight who may also be more economically vulnerable. The perception of unfairness and the qualifier of "model bias" derives from the inability of this group to operate their vehicle in an alternative time frame, not to mention the impact that higher insurance rates can have on groups that tend to be lower income.

Policymakers are then faced with a difficult choice: clearly there are societal benefits to reducing risky driving through the imposition of such a "time of day" rating variable – especially as it relates to individuals who are leaving bars at 2 a.m. Insurers might also want to protect the vulnerable overnight staff from the application of the surcharge. But since use of occupation information in risk classification programs introduces other societal concerns (i.e., a perceived "targeting" of certain population groups), this might not be a viable solution from the regulatory point of view.

A few key themes around bias and fairness have emerged. First, the perception of fairness is enhanced where the relationship between the risk classification elements and the likelihood of outcome are well understood. This acceptance is strengthened where it is clear that the insured individual has the opportunity to use this information and modify their behaviour to realize a benefit. Second, with respect to bias, we note that conflicting opinions around the use of a risk classification element generally position one social benefit against another.

In the real world, these types of issues are frequent and create ongoing challenges for consumers, insurers, and policymakers. These issues also span a variety of financial risk transfer situations, such as:

a) Use of postal code in auto insurance rating

The use of territory in auto insurance rating is controversial, based upon concerns that the practice is an enabler of discriminatory practices where minority groups tend to reside in specific regions. From an actuarial perspective, use of territory in rating has had a long-standing importance based upon the statistical significance of postal codes to predict claim outcomes and brings a clear linkage to road engineering and traffic density that influence driving conditions in an area.

b) Genetic predispositions to health conditions and life insurance eligibility

The potential use of genetic testing in life insurance underwriting has raised a number of issues around the privacy of genetic data and the concern that a known predisposition to a-yet-undiagnosed genetic condition could result in unfair discrimination in other areas.

c) Use of financial responsibility (e.g., credit scores) in homeowners rate-making

Although actuarial studies have demonstrated that these scores have predictive value, it is not easily possible to establish a direct cause-and-effect relationship between an individual's history of financial responsibility and their exposure to property claims. The concern raised by the public is that the use of these scores will lead to higher insurance rates for individuals and socio-economic groups experiencing financial distress.

Since the use of these risk classification elements promotes a sound financial risk transfer system, improved risk reduction incentives, and ultimately a healthier, safer society, we believe that the optimal solution is to permit the most refined risk classification system possible. The question then becomes how to address any resultant social inequities. The two alternatives would be to:

a) Prohibit the use of big data within risk classification if it could potentially create social inequities; or
b) Allow the use of big data in these instances, within the bounds of what is legally permitted, and address any social inequities through other means.

The CIA believes that an efficient system designed to benefit society as a whole depends on the second option.

Suppose that the use of big data determines that a specific region of Canada is much more susceptible to higher property insurance claims due to a number of environmental factors. This area also happens to be predominantly occupied by a visible minority group. There may be a temptation to restrict the use of big data so as to not penalize residents of this area, and, by extension, penalize members of this minority group. Our preferred approach would be to allow the risk classification to proceed using the data available, and if there are perceived social inequities that result, these could be dealt with through public policy measures, provided that the resulting classification system would not be in conflict with legislation.

This should not be construed to mean that insurers should be allowed to cross already established lines; for example, we would not support the explicit use of race or sexual orientation as a rating factor. However, in other cases – an example being the use of age in the determination of life insurance rates – we would support the continued use of established scientific principles. Should there be an external concern about this being seen as discrimination based on age, then the forum for addressing that is through public policy measures, not through restrictions on the use of the underlying data. This is not meant to imply that insurance product design needs to be completely detached from public policy concerns; rather, the use of big data should be expected to comply with regulatory requirements, which also helps to ensure a level playing field for all participants.

It should also be noted that the use of big data may also have the effect of reducing social inequities. For example, insight about an individual's driving habits from vehicle telematics data would reduce the need to rely on generalizations drawn from that individual's age or location of residence.

Call to action

Big data-derived risk classification factors serve an important purpose that protects the underlying mechanisms of an insurance market, and we highlight three key points:

1. **The increasing prevalence of big data in pricing and underwriting within private, competitive insurance markets can be a benefit to society provided that appropriate privacy protections are in place.**

2. **The CIA recommends that big-data-derived risk classification factors are not restricted beyond the ethical data collection practices, privacy laws, and information security requirements necessary to protect consumers (see *Canada's Digital Charter in Action: A Plan by Canadians, for Canadians*).**

3. **The risk classification systems used by insurance companies should be actuarially fair; that is, each individual should be charged a cost that reflects that individual's statistical expectation of their exposure to insured claims, according to all data collected about that individual.**

This is not meant to imply that actuarial considerations should be completely disconnected from other pricing decisions. In fact, actuaries can contribute valuable expertise to the discussion of how to address any potential conflicts between actuarial fairness and social fairness, whether that be at the industry level or within broader society. However, prior restraint on the use of big data is not the appropriate means by which to address these issues. As well, increased uncertainty creates a higher level of risk, which in turn has the potential to increase insurance costs.

The CIA and Canada's actuaries support using big data in risk classification, with the right protections in place, to help reduce risk and establish insurance costs that better reflect the underlying risk.

Sources

American Academy of Actuaries. 2011. *On Risk Classification*.

American Academy of Actuaries. 2014. *Risk Classification Statement of Principles*.

Arrieta, Alejandro Barreido et al. 2020. *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*.

Brown, Robert L. 2010. *The Power of the Collective; The Death of the Collective*.

Brown, Robert L. 1988. *The Canadian Charter of Rights and Freedoms Its Effect on the Canadian Automobile Insurance Industry*. Proceedings of the Casualty Actuarial Society, Volume LXXV.

Canadian Institute of Actuaries. 2011. *Guidance on Fairness Opinions Required Under the Insurance Companies Act Pursuant to Bill C-57 (2005)*.

Canadian Institute of Actuaries. 2014. *Statement on Genetic Testing and Insurance*.

Hao, MingJie, Tapadar, Pradip, Thomas, R. Guy. 2015. *Loss coverage in insurance markets: why adverse selection is not always a bad thing*. International Actuarial Association Colloquium, 7-10 June 2015, Oslo, Norway.

Innovation, Science and Economic Development Canada. 2019. *Canada's Digital Charter in Action: A Plan by Canadians, for Canadians*.

Michael, Liz et al. 2012. *Fairness in Insurance Pricing*. Results of Consumer Research Conducted by the Working Party (GIRO). Institute and Faculty of Actuaries.

Michael, Liz, Ian Hughes, and Andy Goldby. 2013. *Fairness in Insurance Pricing*. Discrimination Working Party GIRO 2021 Presentation plus Update. Institute and Faculty of Actuaries.

National Association of Insurance Commissioners (NAIC) Special Committee on Race and Insurance. 2021. *Principles for Data Collection*.

Office of the Superintendent of Financial Institutions. 2020. *Developing Financial Sector Resilience in a Digital World*.

PwC. 2019. *A practical guide to Responsible Artificial Intelligence (AI)*.

Society of Actuaries. 1992. *Principles of Actuarial Science*. Transactions of the Society of Actuaries, Volume 44

Society of Actuaries. 2019. *Ethical Use of Artificial Intelligence for Actuaries*.

Spindler, Christian and Christian Hugo Hoffman. 2019. *Data Logistics and AI in Insurance Risk Management*. International Data Spaces Association.

Swiss Re. 2011. *Fair Risk Assessment in Life and Health Insurance*.

TD. 2019. *Responsible AI in Financial Services*.

Trowbridge, Charles L. 1989. *Fundamental Concepts of Actuarial Science*.

CIA policy statements address topics of importance to Canadians beyond our usual actuarial practice but where actuarial expertise can contribute to public discourse. Actuaries with diverse backgrounds and views participated in drafting this statement, and all CIA members were invited to provide input to ensure that the statement is supported by a reasonable degree of consensus from our membership.

The CIA would like to thank the members who worked on the development of this statement:

- Emile Elefteriadis, FCIA
- Matthew Buchalter, FCIA
- Christopher Cooney, FCIA
- Blake Hill, FCIA

Canadian Institute of Actuaries
Institut canadien des actuaires

Ottawa, ON K1R 7X7

Canadian Institute of Actuaries
360 Albert Street, Suite 1740
Ottawa, ON K1R 7X7
613-236-8196
head.office@cia-ica.ca
cia-ica.ca
seeingbeyondrisk.ca

The Canadian Institute of Actuaries (CIA) is the qualifying and governing body of the actuarial profession in Canada. We develop and uphold rigorous standards, share our risk management expertise, and advance actuarial science to improve lives in Canada and around the world. Our more than 6,000 members apply their knowledge of math, statistics, data analytics, and business in providing services and advice of the highest quality to help Canadian people and organizations face the future with confidence.