Second Prize Winner

In Praise of Actuarial Judgment: The Dangers of Relying on Historical Data

Sam Gutterman, FSA, CERA, FCAS, MAAA, FCA, HonFIA

Any views and ideas expressed in the essays are the author's alone and may not reflect the views and ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, nor the author's employer.

The thesis of this essay is that over-reliance on looking backward, whether on mortality, morbidity, other actuarial practice-related information/data, or artificial intelligence (AI) training data or information, can be dangerous and can lead to sub-optimal findings and recommendations. Despite the broad potential of AI, supplementary actuarial judgment is necessary for the sound application of AI-generated modeling or findings in an actuarial application.

An initial AI Safety Summit in November 2023 produced the Bletchley Agreement. This resulted in a declaration, signed by the European Union and twenty-eight countries, including China, Saudi Arabia, and the United States, which called on signatories to ensure that AI is "human-centric, trustworthy and responsible." I will focus on all three desirable of these aspects.

AI is about to become even more ubiquitous – with overwhelming attention, it will be the dominant theme of numerous conferences and efforts worldwide in all sorts of applications, as well as a primary driver of a booming stock market.

But before I explore the main theme of this paper, it is worthwhile to address the definition of AI, of which several have been put forth. One that I have seen is that it seeks to mimic human intelligence by inferring patterns and connections in data that are not easily discernible to humans. AI systems can certainly analyze large amounts of data faster than we can by ourselves. They can also develop predictions about future outcomes based on historical information.

Given this definition, let's contrast this with actuarial judgment. But there are also similarities. Pattern recognition is common to both and is an underlying contribution to the effectiveness of many actuaries. But that is only part of actuarial analysis, which can provide timely identification of significant problems or issues (which may depend on the application), as well as quantify the effects of key risks/opportunities and approaches to address them.

Too often, actuaries have been characterized as relying solely on historical experience and information – but this can also be a criticism of AI. Both rely on historical experience or information (i.e., referred to as training data in AI), based on a deep analysis of patterns in available data applied in a systematic and algorithmic manner. What may be different from the AI repertoire is that a skilled actuary also applies judgment to the problem at hand.

AI algorithms can develop and present predictions by correlating information and applying patterns from past events or adjustments based on historical data and past learnings. AI models can process vast amounts of historical data quickly, organizing them by rules and using labels ultimately provided by humans, possibly containing an inherent element of bias, although an AI (human) supervisor can train the AI model to identify and thus ignore this bias, at least to the extent provided by the supervisor.

But some AI models can be black boxes — they may not adequately explain to users how they arrived at their conclusions; nor why their insights should or should not be relied upon. If they are wrong, a danger is that AI systems will undermine the trust given to them and in some cases social trust, by leading their users astray.

Although constraints and considerations could be built into an AI model's algorithm, there are certain factors, such as race or ethnicity, that would be unacceptable to use in a function such as insurance pricing but may be acceptable for use in a public policymaking process.

THE DATA, THE PROBLEM, AND THE CONTEXT

Identification and assessment of reliable and relevant data/information input are crucial actuarial functions. A great deal of experience analysis units inside an insurance company, for example, are involved in this sometimes time-consuming, tedious, yet complex process. Hidden problems in experience data exist, just as they can affect the data used in training an AI model. Once reliable and relevant data/information are obtained, analysis can begin.

AI can't by itself ask all the right questions and exhibit skepticism, although they can be applied to a preconceived problem or issue. AI may not, by itself, be able to identify the quality of or bias in the data obtained or to appropriately weight data/information obtained from different sources. It could, by applying an algorithmic rule, assign a degree of credibility to sets of data based on size; but it may not be able to assess reliability or relevance, which may only be applicable to the issue being addressed.

For example, a common problem is dealing with an outlier, which may prove to be an early warning indicator or an indication of faulty data that might be safely ignored. I asked ChatGPT for accepted approaches to assess outliers. What it 'told' me is a decent start in analysis: it depends on the context and purpose of the analysis. It then proceeded to provide some common approaches to use:

1. "Understanding the cause, that is, whether due to measurement error or natural variation;
2. data cleaning;
3. transformation or normalization;
4. statistical methods such as regression or median-based measures;
5. segregation if genuine or imputation if erroneous, that is, analyzing them separately or substituting values surrounding them;
6. modeling; and
7. reporting (transparently) and justifying decisions made in light of the outliers."

It then indicated that the choice of the method used "should be guided by a combination of statistical rigor, domain knowledge, and the specific goals of your analysis." This is certainly (and surprisingly) good advice. But it leaves the choice as to which approach to use to the analyst (who has domain knowledge) – which is the role of actuarial judgment. Also, it suggested that the person (people) who is overseeing the use of an AI tool should ensure that the data/information used is not misinterpreted, is contextually applicable, and whether one or multiple approaches should be applied sequentially or at the same time.

This reminds me somewhat about the time that I over-fitted a set of data. I was so uber-focused on past experience, that is, I slavishly followed patterns, trends, and covariances derived from the past, when subsequent changes in underlying conditions or populations over the period studied would have led me to quite different conclusions. It is easy to lose sight of the objectives of an analysis, which should include an assessment of future risks that should have been anticipated and acting dynamically in response.

On the other hand, the use of the output of an AI model might encourage the user to think more in-depth regarding where to be skeptical, what questions to ask, and what explicit assumptions to make.

The context is important in many actuarial analyses, e.g., in regular monitoring, pricing, valuation, liability analysis, and risk management. Different analytical approaches may be appropriate for each. One model, whether or not an AI one, cannot apply to all applications. The user can provide an AI model with a description of what to do, and some hints (leading language) about how to approach the analysis. Just as the use of a computer is quite a bit more efficient than manually prepared spreadsheets, actuarial-specific software that is regularly used can be even more efficient and useful. Incorporating AI features can, for example, create alternative pre-defined or AI-generated scenarios faster than a new set of scenario-specific assumptions in non-AI software that would have to be created from scratch each time applied.

Many actuarial exercises are predictable – conducted once a year, quarter, or month when an analysis of new data is combined with previous data to form new patterns to use as input into a pricing, funding, or liability calculation. However, that said, I would hate to present to a CEO results that have not been properly supervised, governed (e.g., through a peer review or multi-disciplinary governance committee), and validated.

CHANGES IN CONDITIONS

Training a model or its assumptions on historical data/information can be quite valuable. Unless, of course, the data/information used directly or for training purposes is not relevant to what that experience will be applied to. Certainly, the job of an actuary is often to assess or project the future, but it may be just as if not more important to select the most appropriate historical experience to serve as its basis and to determine the types of adjustments are needed to derive a reasonable estimate of future experience. This places a premium on estimating the extent future risk factors will be different from those of the past.

Because actuarial analysis tends to focus on such risks, I asked ChatGPT what the most significant upcoming mortality risks are. It indicated:

1. "Health-related risks.
    a. Non-communicable diseases: Conditions such as heart disease, cancer, and diabetes continue to be leading causes of death globally. Heart disease remains the top killer, followed by cancer and diabetes-related complications.
    b. Pandemics and infectious diseases: The threat of global pandemics, including influenza and emerging high-threat pathogens like Ebola remains significant. The COVID-19 pandemic has highlighted the devastating impact of infectious diseases on global mortality.
    c. Antimicrobial resistance: This is a growing concern as it can lead to common infections becoming untreatable, resulting in higher mortality rates from previously manageable diseases.
2. Climate-related risks.
    a. Extreme weather events: Increasingly severe weather events such as hurricanes, floods, and heatwaves pose significant mortality risks. These events are expected to become more frequent and intense due to climate change.

   b. Biodiversity loss and ecosystem collapse: The degradation of ecosystems can disrupt food supplies, increase the spread of diseases, and lead to loss of livelihoods, which in turn affect mortality rates.
   c. Critical changes to Earth systems: Climate tipping points, such as the melting of polar ice caps and shifts in ocean currents, can have far-reaching and potentially catastrophic impacts on human health and survival.
3. Geopolitical Risks.
   a. Conflict and instability: War and political instability can lead to direct mortality from violence, as well as indirect mortality through the destruction of healthcare infrastructure, displacement, and famine.
   b. Misinformation and disinformation: The spread of false information, particularly regarding health and safety, can exacerbate public health crises and hinder effective responses to emergencies."

When I further queried it regarding which of these risks are most relevant to high-income countries, it came back with:

1. "Non-communicable diseases: heart disease and stroke, cancer, and diabetes.
2. Mental health and substance abuse.
3. Climate-related events, including extreme weather events, and air pollution.
4. Pandemics and infectious diseases: influenza and respiratory diseases, and antimicrobial resistance.
5. Lifestyle-related diseases: obesity and related complications, smoking, and alcohol use."

I could have then proceeded to ask, for each of these risks for further recent research to determine the likelihood that these risks will occur or become more severe, and what age groups they might affect most.

After a further inquiry regarding the best process used to project mortality improvement, it concluded that "Projecting future mortality improvements is a complex but manageable task involving careful analysis of historical data, health trends, and potential future changes in medical care and public health policies. By using robust statistical models and considering a range of scenarios, more accurate projections can be developed to inform public health planning and policymaking." Certainly, good advice for actuaries developing mortality improvement assumptions.

This suggests that, with further inquiries, AI can provide valuable input and assistance, listing the basic steps that an actuary could follow. However, the execution of each step should be handled or overseen by a well-trained and experienced actuary to complete an analysis or other project.

THE GUT CHECK

Sometimes a check for reasonableness is referred to as a gut check or intuition. Are the results (and intermediate findings) reasonable? In practice, these findings are often professionally peer-reviewed to determine if the relevant factors have been considered, appropriate processes followed, and are properly documented. These parts of model governance need to be incorporated and assessed in the use of any algorithmic approach. A review of the underlying data or information could focus on the data/information that the AI was trained on, if obtainable, with the results being thoroughly validated and checked for reasonableness.

I draw an analogy to predictive keyboarding – my word processor pre-shadows my typing in many cases to help me become more efficient, but its suggestions are wrong almost as often as they are right. At the same time, it also can be distracting. Similarly, if an actuary depends too much on a trained algorithm without proper supervision or reasonableness checking, the risk is that the findings, conclusions, or recommendations could be wrong or lead the actuary astray.

CONCLUSIONS

AI is a hugely important technical development, representing a potentially valuable tool that may be able to enhance the practice of an actuary. If mastered and properly directed, it can serve multiple purposes. Yet its algorithmic findings cannot replace the key functions of an actuary and may not even identify the initiating problem. In addition, it is unlikely to replace the identification and assessment of the quality of data/information source, the choice of the most relevant trajectory of analysis, and the assessment of the reasonableness of findings, especially in the development and use of scenarios. Just like other information development and modeling, it can enhance the efficiency of the processes followed. Nevertheless, the prospective nature of actuarial work demands checks and reviews throughout our processes. Without proper 'supervision,' over-reliance on an AI model can lead to failure and added risk.

POSTSCRIPT

After I finished this essay, I thought it appropriate to obtain a second opinion regarding the reliability of an AI model. To do so, I asked ChatGPT. Its conclusion was eerily similar to what I have discussed in this essay:

"Factors influencing AI model reliability:

1. Quality of training data: Diversity and representation. The training data must be diverse and representative of the real-world scenarios the model will encounter. Bias in the data can lead to unreliable and unfair outcomes, and quantity of data. More data generally leads to better performance, as the model can learn from a broader range of examples.
2. Model design and complexity: Algorithm choice: Different algorithms have different strengths and weaknesses. The choice of algorithm impacts the model's ability to generalize from the training data to new data. Overfitting/underfitting: A reliable model strikes a balance between overfitting (too closely matching the training data, reducing generalizability) and underfitting (failing to capture patterns in the training data).
3. Testing and validation: Cross-validation: Techniques like cross-validation help ensure the model performs well on unseen data by splitting the data into training and testing sets. Performance metrics: Metrics such as accuracy, precision, recall, F1 score, and area under the curve provide insights into the model's performance and reliability.
4. Robustness to changes: Adaptability: A reliable AI model can handle variations and changes in the input data over time. Stress testing: Models should be tested under a variety of conditions to ensure they are robust and reliable in different scenarios.
5. Transparency and explainability: Interpretability: Models that are easier to interpret and explain are generally more reliable because their decisions can be understood and validated by humans. Bias and fairness audits: Regular audits for bias and fairness help ensure the model's reliability and ethical integrity.

Reality in Practice

1. Healthcare: AI models in healthcare must be highly reliable as they can directly impact patient outcomes. Rigorous testing, validation, and regulatory compliance are critical. Example: AI models for diagnosing diseases from medical images have shown high accuracy in controlled studies, but real-world reliability requires extensive validation and oversight.

2. Finance: In finance, AI models are used for credit scoring, fraud detection, and trading. Reliability is crucial to avoid significant financial risks. Example: AI models for credit scoring must be continuously monitored to ensure they do not introduce bias against certain groups and remain accurate as economic conditions change.
3. Autonomous vehicles: Autonomous vehicle AI systems must reliably interpret and respond to a wide range of driving conditions to ensure safety. Example: Despite advancements, the reliability of autonomous driving systems is still a concern, with ongoing development needed to handle edge cases and improve overall safety.

Limitations and challenges

1. Data privacy: Ensuring data privacy while using large datasets for training can be challenging.
2. Bias and fairness: Addressing inherent biases in training data is an ongoing challenge, as biased models can lead to unfair and unreliable outcomes.
3. Dynamic environments: AI models may struggle with rapidly changing environments unless they are regularly updated with new data and retrained.

Conclusion

The reliability of an AI model is contingent upon careful design, thorough testing, and continuous monitoring. While AI models can achieve high levels of accuracy and reliability, it is crucial to address limitations such as bias, data quality, and adaptability to changing conditions to ensure they remain trustworthy and effective in real-world applications."

Now that I read what ChatGPT produced regarding AI models and eliminated the jargon, it looks similar to what I would say about actuarial modeling and practice, with the addition of the use of actuarial judgment and a profession-wide self-governance (in most cases) process.

* * * * *

Sam Gutterman, FSA, CERA, FCAS, MAAA, FCA, HonFIA is a consulting actuary. He can be reached at sam.gutterman1@gmail.com.