# Faster close reserving

## Using numerical diagnostics to make faster and smarter decisions on roll-forward

**August 2026**

**Oli Grossman**
*oliver.grossman@lcp.com*

**Ed Harrison**
*ed.harrison@lcp.com*

**Charlie Stone**
*charlie.stone@lcp.com*

*Being able to quickly identify which classes are at risk of reserve deterioration or improvement, and why, is a key challenge for reserving actuaries, especially on roll-forward, when internal company timelines are often tight. Reserving actuaries typically need to review a large number of reserving diagnostic charts for each class in order to identify risks of future deterioration. This makes the review stage of the actuarial control cycle time-consuming and inefficient relative to the other more automated stages of a typical reserving or roll-forward process.*

*Our paper demonstrates that we can summarise important diagnostic information from graphical reviews into numerical alignment scores, providing a more efficient screening tool for actuaries. We further propose a novel machine-learning-led framework for predicting reserve deteriorations, which incorporates these scores into a logistic regression model and uses SHAP analysis to identify which of these diagnostics contributes most to the deterioration risk for each reserving class.*

*When tested on approximately 1,300 classes using the NAIC Schedule P returns, our multi-parameter model correctly identified 82% of classes that went on to experience a reserve deterioration of >5% over the next 3 years, whilst only requiring the reviewing actuary to perform a more detailed analysis on the 63% of classes flagged as being at risk. This provides useful predictive performance and time-saving potential when applied to a multi-company, multi-business-class dataset for a large and mature insurance market.*

*Overall, we set out to help actuaries make faster and smarter decisions during the reserving process, particularly on roll-forward when time is most pressured. Faster decision-making comes from using our model to highlight classes that are at risk of deterioration and screen out those that are not. Smarter decision-making comes from the SHAP analysis clearly highlighting which diagnostics are most concerning for any classes flagged as a deterioration risk.*

## Contents


1. Introduction ..... 3
2. How actuaries typically review the reasonableness of reserves ..... 4
3. Our approach ..... 6
4. Metric alignment scores ..... 6
5. A multiple metric model ..... 15
6. Model validation ..... 28
7. Interpretability and the human in the loop ..... 34
8. A two-way model that captures reserve improvements and deteriorations ..... 37
9. Key conclusions and implications for reserving practitioners ..... 37
10. Limitations ..... 39
11. Recommendations for further work ..... 40
12. References ..... 41


## Disclaimer

*The views expressed in this paper are those of the authors and do not necessarily reflect the views of their employer, clients, the Institute and Faculty of Actuaries, or any other organisation with which they are associated. The analysis and conclusions presented are intended to contribute to discussion and research and should not be relied upon as actuarial, financial, legal or other professional advice.*

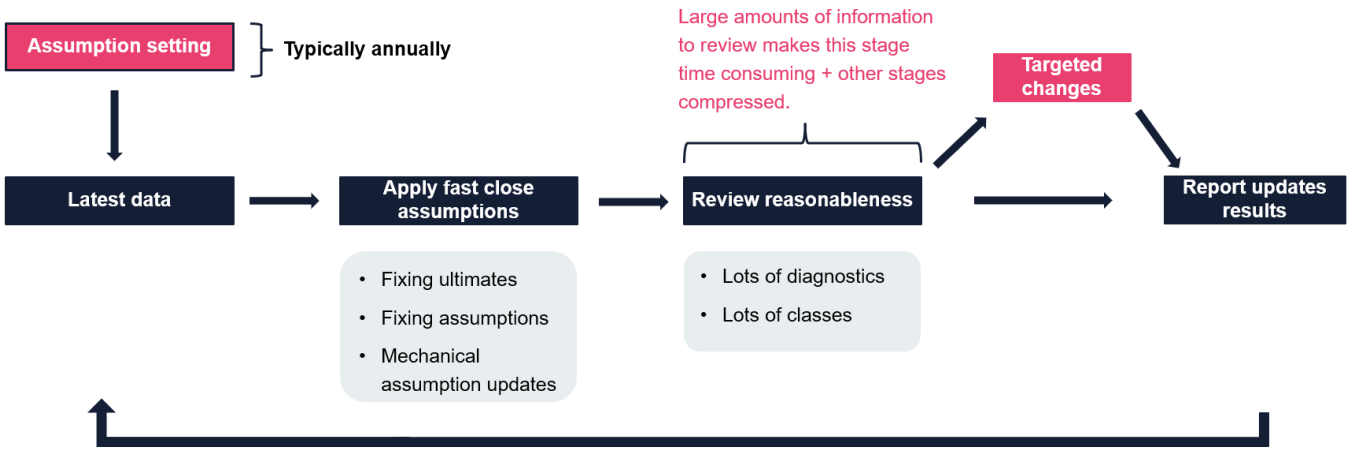
# 1. Introduction

In 2014, the IFoA's Reserving Oversight Committee set out the "Fast Close Process" [1] for reserving. This demonstrated a range of reserving techniques and process steps that allowed for frequent (ie monthly or quarterly) updates to reserving results, without the need for detailed review of assumptions each time.

Wide adoption of this approach across the UK personal lines, commercial lines and London markets has shifted reserving from being a periodic exercise to a regularly updated process. This has allowed for faster feedback into the business, with updates often coming monthly or quarterly instead of half-yearly or annually.

The fast-close process in its simplest form is set out in Figure 1, below.

**Figure 1: Outline of a typical fast close reserving process**




The typical fast close process involves obtaining the latest data, applying reserving assumptions automatically using a series of pre-determined rules, and then reviewing the output for reasonableness before publishing the final results for internal discussion within the company.

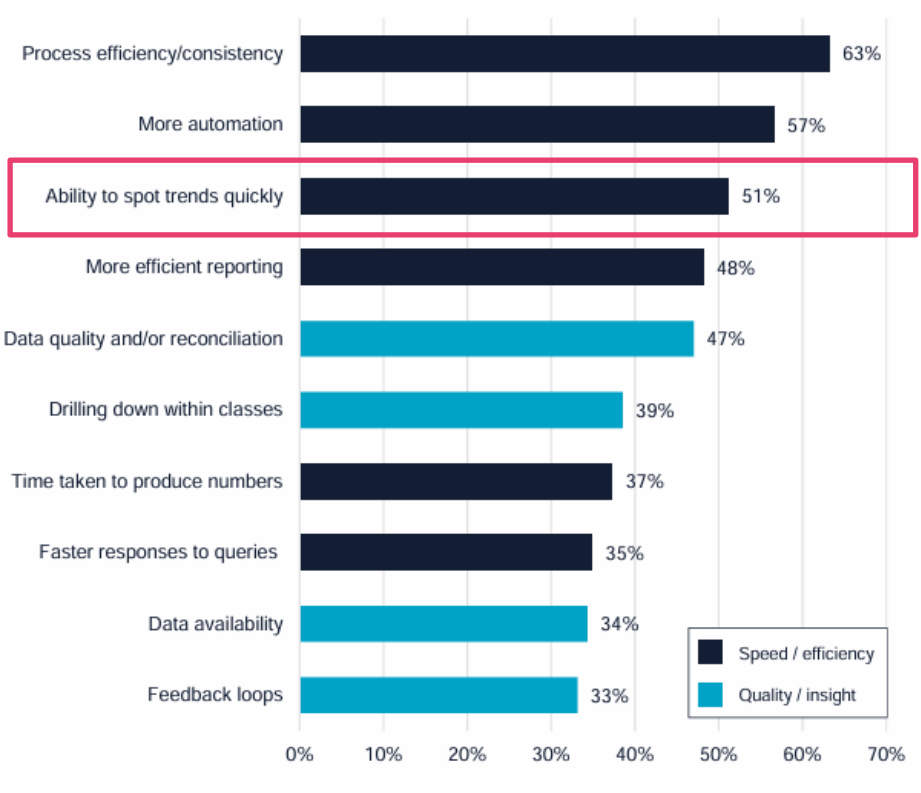
Automation of data processing, application of assumptions and generation of reporting dashboards are now embedded as relatively standard market practice. As a result, the adoption of the fast-close process has shifted the focus for reserving actuaries towards being able to efficiently identify trends of concern and highlight which classes are performing particularly favourably or poorly.

Actuaries use a range of numerical and graphical diagnostics to help spot concerning trends and validate their projections, as explored by Balester and Kirschner [2] and by the IFoA Towards the Optimal Reserving Process Working Party [3]. Section 2 provides further summary of the typical steps involved in current review processes. The key challenge is that many of these diagnostic techniques rely on graphical human review to spot trends, meaning that they must be opened and reviewed by eye across a large number of classes, making the reasonableness checks stage of a reserving process or a roll-forward time-consuming.

Reasonableness reviews and subsequent targeted methodology changes are an efficiency challenge, but particularly so on roll-forward, when company timelines are often tight and actuarial teams have a limited time budget to assess the ongoing reasonableness of their assumptions and results across a large number of classes.

Industry surveys have confirmed that being able to spot trends quickly to assist with reasonableness checking is a high-priority area for improvement for insurers globally as shown in figure 2 from LCP's 2024 global reserving and transformation survey [4]. In the same survey, firms also ranked process efficiency and automation highly, but the reason for this was actually often to create more time for the trend spotting and reasonableness checking stage.

**Figure 2: Key areas for reserving process improvement – LCP 2024 survey of 166 insurers**




## 2. How actuaries typically review the reasonableness of reserves

We see two typical approaches to reviewing reasonableness in the market.

- Diagnostic review
- AvsE review

### 2.1. Diagnostic review

Reserving actuaries typically consider a wide range of diagnostics when reviewing the reasonableness of their assumptions and results.

Metrics frequently used by actuaries as part of a graphical diagnostic review include:

- **Incurred (and paid claims) as percentage of projected ultimate claims:** This is a widely used diagnostic chart, used to assess alignment between current reserving results and actual past experience.
- **Method comparison charts:** These highlight differences and trends between alternative methodologies, to provide an indication of uncertainty in actuarial projections.
- **Comparisons of development patterns based on different averaging periods:** These highlight places where there may be a change in claims development depending on whether the chain ladder model is fitted to lots of historical data, or just the most recent experience.
- **Annualised frequency, average cost or loss ratio charts and implied year-on-year inflation rates:** to identify years where implied year-on-year inflation may seem higher or lower than expectation.

The above list is not intended to be exhaustive, but simply to highlight that there are a wide range of diagnostics that actuaries use in combination to assess reasonableness of reserves.

Manual review of diagnostics is time-consuming and judgement-driven process, and can yield inconsistent results. This is because:

- The diagnostics are typically graphical, and it requires concurrent review of several graphs to validate whether a class appears reasonable or not.
- The graphical nature means that reviewing a class can require examination of several pages of graphs.
- Graphical review is inherently subjective, meaning that two actuaries interpreting the same data are at risk of reaching different conclusions.

## 2.2. Actual vs expected analysis

Actual vs expected analysis (AvsE) is a commonly used tool, that prioritises review of classes of business based on any deviation of actuarial experience compared to that expected at the previous review date. This is integrated into the technical actuarial standards as a requirement, and was explored in detail by the IFoA ROC working party in 2013 [5].

Despite its prevalence, AvsE has a number of limitations that can reduce its effectiveness as a high-level diagnostic, including:

- **Reliance on a flightpath / cashflow pattern:** The "expectation" element of AvsE is usually set with reference to the paid cashflow or incurred development pattern, derived from a chain ladder or similar model. However, the chain ladder may not be used as the predictor of ultimate losses because it is not sufficiently mature or reliable. For example, an IELR based approach might be used instead. Actuaries compensate for this by re-scaling the implied incurred movements or cashflows to match the ultimate claims. However, this scaling can be distorting and does not address the fact that actual emerging experience is being evaluated relative to a development pattern that is not trusted as a predictor of the final ultimate. In some cases, there may not be a development pattern set at all, which then makes AvsE analysis extremely challenging.
- **Accounting for uncertainty:** AvsE movements month-to-month or quarter-to-quarter can be significant, simply as a result of volatility, rather than caused by a specific trend or under/over reserving. This is not captured numerically by the methodology and must be overlaid based on the judgement of an experienced actuary, often with the support of graphical reviews. For example, a large loss movement may be within the bounds of normal volatility, but it will nonetheless show up as a significant AvsE residual.
- **AvsE can be misleading for immature cohorts:** Where a class is relatively immature, both the actual and expected movements may be small. As a result, the absolute AvsE movements may not be flagged for review because they too are small, even where they might be providing a clear indication of an adverse trend with a more significant impact on ultimate claims.
- **Numerical evaluation only covers experience since the last review, rather than considering longer term trends.** This limitation can be avoided by graphical review of longer term AvsE, but this is then a graphical analysis rather than the more efficient tabular consideration.
- **Aggregate AvsE for any class can be distorted by single large movements:** for example a large loss movement on an older cohort can offset evidence of model bias on more recent cohorts – unless a more detailed review of the accident period AvsE is performed.

# 3. Our approach

We set out to make the reasonableness checks stage of reserving processes, particularly on roll-forward, both faster and smarter.

To make the process faster, we seek to:

- Reduce the emphasis on diagnostic review by demonstrating that key actuarial insights normally obtained by graphical review can instead be captured by numerical analysis.
- Show that the results of this type of numerical analysis can be predictive of future reserve deteriorations when several “alignment metrics” (and others) are combined in a multi-parameter model.
- Demonstrate that applying this approach can highlight the classes that an actuary needs to review in more detail and distinguish these from classes that do not require detailed review.

To make the process smarter, we seek to ensure that the underlying drivers of reserve deterioration risk (ie which metrics matter and why) can be recovered from the multi-parameter model at the end, so that the actuary has the same understanding as if they’d reviewed the graphs, but in much less time.

As part of this work, we also look to avoid some of the key limitations of an AvsE analysis, ensuring that if our proposed approach were used in conjunction with AvsE, it would be complementary.

The next four sections of this paper summarise our analysis, and are set out as follows:

- **Section 4: Metric alignment scores.** We set out a formula to calculate what we describe as the incurred (paid) alignment score – a metric that captures the key features actuaries look for in an incurred (paid) as percentage of ultimate diagnostic chart. We demonstrate that this formula is predictive of reserve deteriorations on Schedule P data.
- **Section 5: A framework that combines multiple metrics.** We extend our analysis to consider multiple alignment metrics in a machine learning based framework, together with other information relevant to predicting reserve deteriorations. We combine these into a five parameter logistic regression model and analyse in detail its ability to predict reserve deteriorations.
- **Section 6: Model validation:** We perform a series of validation tests to demonstrate the reasonable performance of this model in a broad range of conditions.
- **Section 7: Interpretability and the human in the loop:** We demonstrate how SHAP analysis can be used to highlight which model parameters drive the deterioration risk for each reserving class the model is applied to, thus ensuring the model remains explainable and the drivers of deterioration risk remain clear to a reviewing actuary.

# 4. Metric alignment scores

## 4.1. Overview

In this section, we derive the metric alignment score that captures the key features of common reserving diagnostics. The purpose of the metric alignment score is to capture numerically the key information from analyses that actuaries have historically performed graphically. In the case of incurred (or paid) to ultimate ratios, this is designed to identify the extent to which the latest incurred (or paid) to ultimate ratio is consistent with past experience – a key feature that actuaries would look for when reviewing graphically, and considered to be a key determinant of the reasonableness of reserving assumptions.

Although we focus on incurred and paid to ultimate ratio metrics in this paper, the alignment score is KPI agnostic and can be used to evaluate any triangle-based diagnostic where you are looking for consistency between current

projections and actual historical development experience. As such, it could easily be extended to claim numbers, average cost and claim settlement projections.

After defining this metric, we go on to test its performance as a predictor of reserve deteriorations based on the Schedule P industry data. Further details regarding the Schedule P data are set out in section 5.2.

## 4.2. Incurred and paid to ultimate ratios

Incurred (or paid) to ultimate ratios are calculated based on the following formula:

$$M_{i,j} = \frac{C_{i,j}^{\text{inc}}}{U_i}$$

where:

- $C_{i,j}^{\text{inc}}$ are the cumulative incurred (or paid) claims for accident period $i$ at the end of development period $j$.
- $U_i$ is the projected ultimate claims cost for accident period $i$, available at the valuation date.

Incurred (or paid) to ultimate ratios are typically plotted and reviewed graphically by reserving actuaries. They are considered to be effective high-level indicators of whether a reserving class is over or under reserved. Actuaries look for consistency between the current incurred (or paid) to ultimate position of each cohort and observed historical experience at the same development point.

Figures 3 and 4 below provide examples of how this diagnostic is commonly displayed and interpreted.

**Figures 3 (left) & 4 (right): Example incurred to ultimate charts**

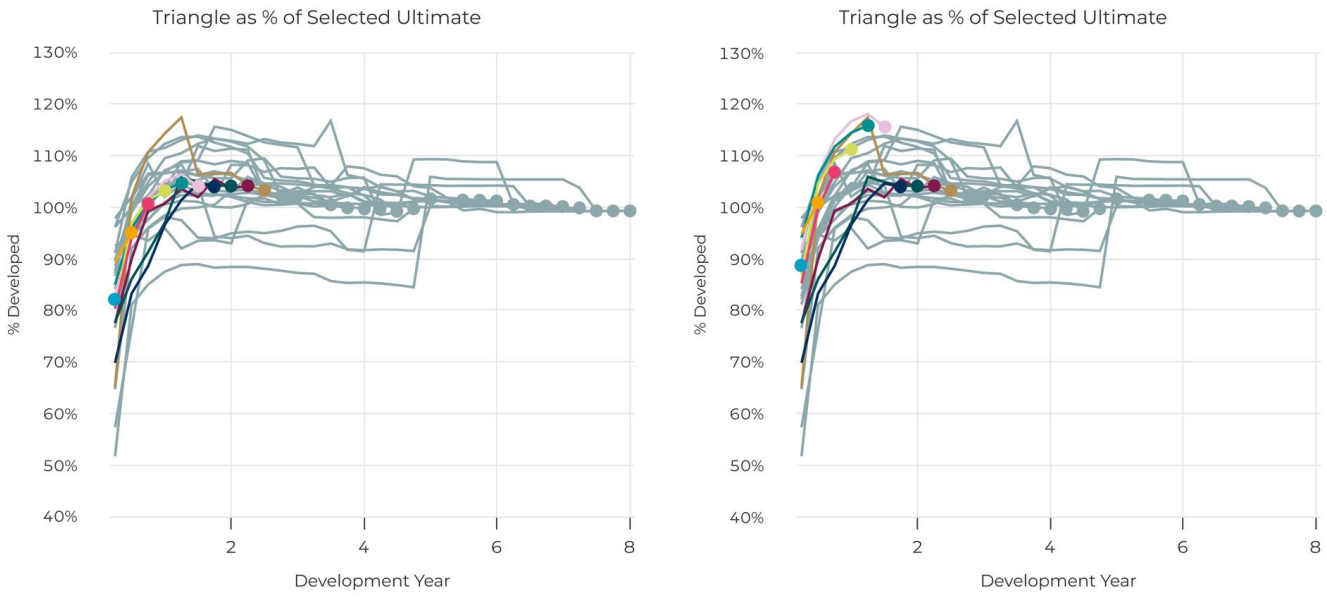



Figure 3 shows that the current incurred position of each cohort is consistent with observed past experience. In essence, this shows that the current IBNR allowance is consistent with typical historical requirements at the same point in time.

Figure 4 shows the same underlying incurred data as figure 3, but on a more optimistic reserving basis on more recent cohorts. This means the incurred as % of ultimate is higher with those recent cohorts having fewer reserves than actual past experience would suggest have been required. This would be flagged as a risk by an experienced

reviewing actuary and would prompt further investigation. It is this type of case that the proposed numerical metric is intended to flag.

The above analysis is a part of normal market review practice for many actuarial teams but involves (1) a relatively detailed graphical review and (2) sufficient experience to interpret the charts for each class - looking for evidence of alignment between the latest data position (the dots on the chart) and historical experience (the lines).

## 4.3. Derivation of the incurred (or paid) alignment score

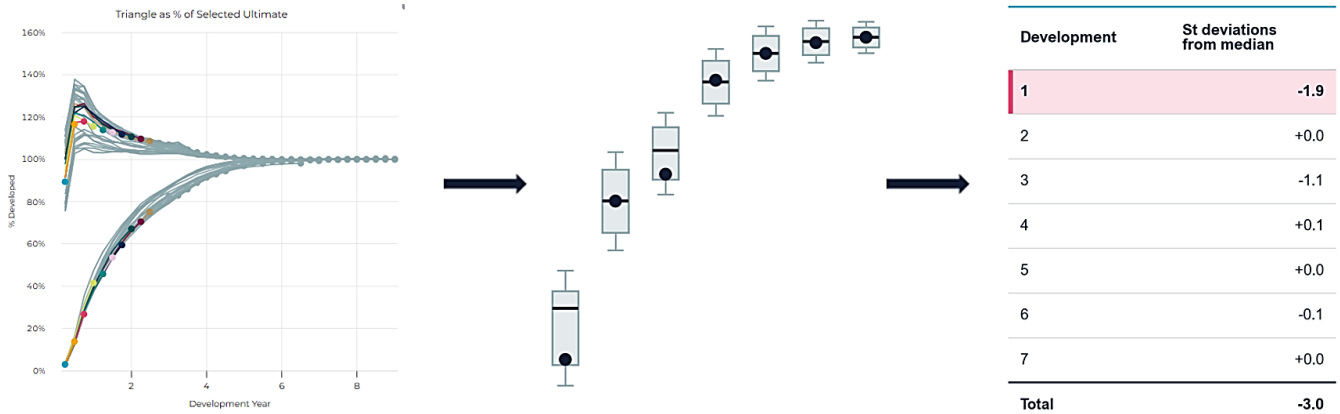
We adopted the following process to calculate a metric alignment score, which captures the key diagnostic features that actuaries look for in these incurred (paid) to ultimate ratio charts. In essence it provides a numerical score that highlights whether data is reserved consistently with past experience, above it, or below it. The magnitude of the score highlights how aligned or misaligned the ultimates are with that experience.

In simple terms our approach is to:

1. Calculate incurred (or paid) to ultimate triangles.
2. Calculate the standard deviation of the data in each development column of the triangle (excluding the latest accident period).
3. For each column, calculate the number of standard deviations that the latest accident period (the one on the lead diagonal for that column) is from the mean value.
4. Apply a score to each column based on how many standard deviations it is from the mean, capping individual outliers at a maximum of +/- 5 standard deviations from the mean.
5. Combine the individual scores, weighting each according to the number of data points in each development column, so that scores for immature cohorts are increased, and more mature ones receive less weight.
6. Scale the composite score for the number of accident periods in the data, to avoid large triangles automatically leading to large scores.
7. Cap and collar the final score between +/-10 to avoid outlier results.

The broad process is summarised in figure 5 below. Although we refer to accident periods above, this approach should be equally valid for underwriting or reporting period cohorts.

**Figure 5: outline process for converting incurred to ultimate ratio into a numerical score**




| Development | St deviations from median |
|---|---|
| 1 | -1.9 |
| 2 | +0.0 |
| 3 | -1.1 |
| 4 | +0.1 |
| 5 | +0.0 |
| 6 | -0.1 |
| 7 | +0.0 |
| **Total** | **-3.0** |

In steps 4 – 7, we make specific adjustments to allow for:

- **Variations early in the triangle being of more significance than those in the tail** – because that's where there are typically more reserves. We allow for this by weighting the score by the number of cohorts included in the column.
- **Capping / collaring of extreme outliers**, eg where a single data point is multiple standard deviations from the mean and would otherwise distort the overall alignment score. We cap each data point at a maximum of 5 standard deviations from the mean.
- **Capping of the overall score between -10 and 10,** again to limit the effect of extreme outliers on the final score.

The alignment score metric *A(M)* is defined as follows for KPI (M), where M is typically paid or incurred to ultimate ratio:

$$A(M) = \max\left[-10, \min\left(10, \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}} n_i \max\left(-5, \min\left(5, \frac{M_{i,j_i^*} - \mu_i}{\sigma_i}\right)\right)\right)\right]$$

Where:

- The latest observed development period j for accident period $i$ (ie the point on the lead diagonal) is defined as:

$$j_i^* = \max\{j : \exists\, M_{i,j}\}$$

- The peer set $\mathcal{P}_i$ contains the other accident periods in the same development period (ie the relevant column of % developed in the triangle, excluding the accident period on the lead diagonal) and is defined as:

$$\mathcal{P}_i = \left\{r \neq i : \exists\, M_{r,j_i^*}\right\}$$

$$n_i = |\mathcal{P}_i|$$

  - The peer set (development column) mean and sample standard deviation are respectively:

$$\mu_i = \frac{1}{n_i}\sum_{r\in\mathcal{P}_i} M_{r,j_i^*}$$

$$\sigma_i = \sqrt{\frac{1}{n_i - 1}\sum_{r\in\mathcal{P}_i}\left(M_{r,j_i^*} - \mu_i\right)^2}$$

  - $\mathcal{C}$ is the set of accident periods where the score is defined:

$$\mathcal{C} = \{i : n_i \geq 2, \sigma_i > 0\}$$

This formula identifies places where, in aggregate, the incurred to ultimate ratio is high or low compared to actual past experience, focusing the review on a single numerical result. The output of the formula is a numerical score for the reserving class between -10 and +10 highlighting a relative degree of prudence or optimism. Scores can be viewed at the cohort level or as a single score for each triangle.

## 4.4. Applying alignment metrics to a real world dataset

We tested our alignment metrics on actual historical incurred and paid to ultimate data from the Schedule P returns.

We calculated incurred and paid as % of ultimate triangles for each company and business class combination, for two historical periods, as per the table below.

### Table 1: summary of assessment methodology

| Metrics fitted to the following data period | Reserve deteriorations assessed against | Incurred (paid) triangles in dataset |
|---|---|---|
| 2002 - 2008 | 2011 | 5,619 |
| 2013 - 2019 | 2022 | |

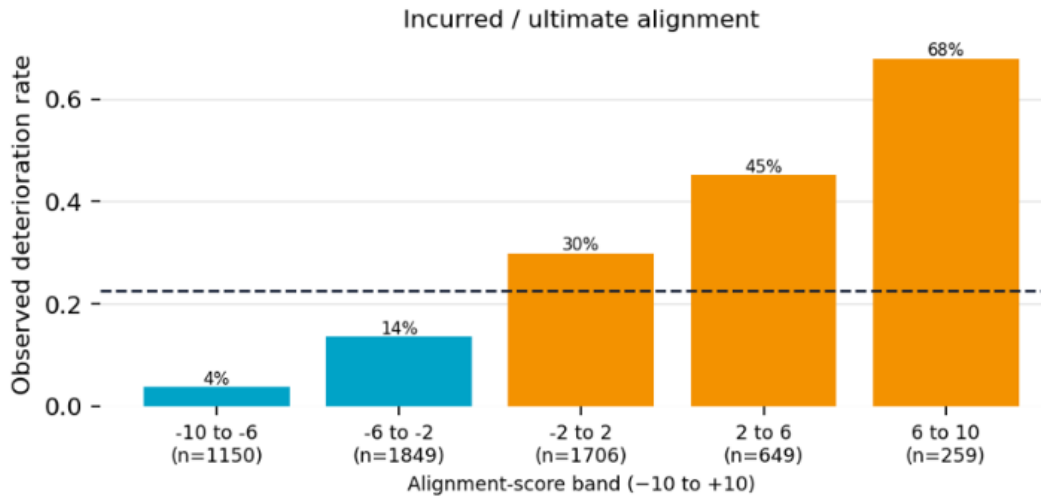
We used a 7 accident year x 7 development year triangle to calculate the alignment score for each company, class and time period – 5,619 incurred triangles in total (and the same number of paid triangles). There are however 6 triangles where this metric cannot be calculated, hence we assess 5,613 triangles.

We then reviewed the ultimate claims data for the same company and class after a further 3 years to identify if that class had deteriorated. We counted a class as having deteriorated if the ultimate claims had increased by 5% or more over the 3 year time period. We outline the data used further in section 5.2.

## 4.5. Results – incurred alignment score

Figure 6 (below) compares the proportion of classes where a reserve deterioration was actually observed vs the banded incurred alignment score (-10 to +10). The dashed line shows the overall deterioration rate for the data (23%). Orange columns represent bands with a deterioration rate above the overall average; blue columns represent bands below it.

### Figure 6: proportion of classes deteriorating, based on incurred alignment score across the 5,613 triangles




We can see there is a clear correlation between increasing incurred alignment score and observed reserve deteriorations in the data. Reserving classes with a higher score have a significantly greater probability of experiencing a deterioration over a 3-year time horizon than those with a low score. This analysis suggests that incurred alignment scores may be predictive of reserve deteriorations and offer a more efficient alternative to purely relying on graphical analysis of the same data, class by class.

Using this data, we can derive an indicative calibration table for application to comparable datasets to flag their risk of reserve deterioration. A suggested calibration is set out in table 2 below that would allow actuaries to apply the incurred alignment formula to their data and then evaluate each class for reserve deterioration risk.

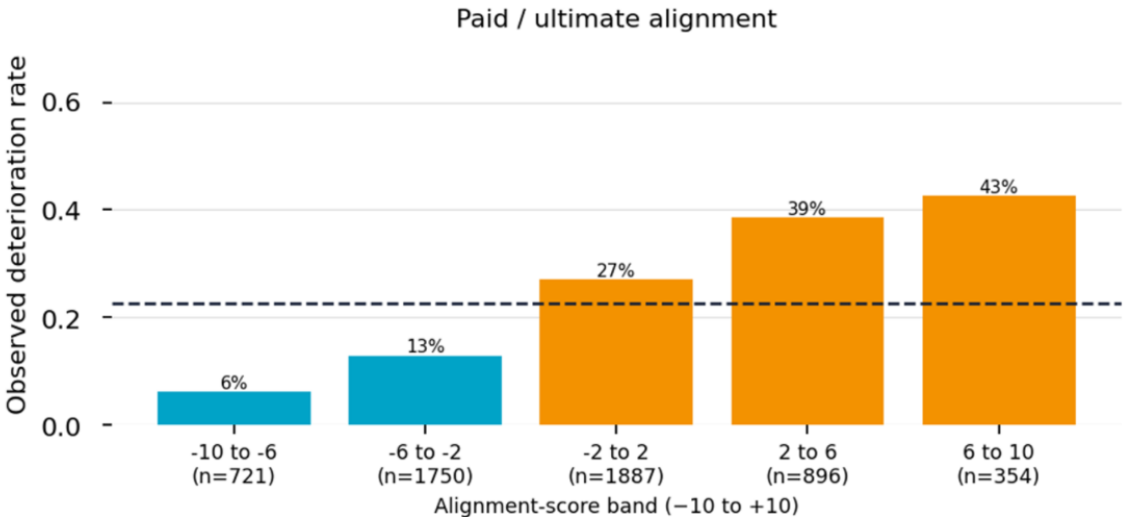
**Table 2: interpretation of results from the incurred alignment score**

| Incurred alignment score | What you would expect to see in a diagnostic graph | Interpretation |
|---|---|---|
| A < -6 | Most cohorts sit well below past data | Likely prudent |
| -6 ≤ A < -2 | Some early cohorts below past data and / or many cohorts towards bottom of past data. | Possibly prudent |
| -2 ≤ A < 2 | Cohorts overall well aligned with past data | Approximately best estimate (eg figure 3) |
| 2 ≤ A < +6 | Some early cohorts above past data and / or many cohorts towards top past data. | Possibly optimistic – risk of deterioration |
| 6 ≤ A | Most cohorts sit well above past data | Likely optimistic – high risk of deterioration (eg figure 4) |

## 4.6. Results – paid alignment score

We performed similar analysis to calculate the paid alignment score, substituting the incurred metric for paid and using the same data groupings. The results are shown in figure 7 below.

**Figure 7: proportion of classes deteriorating, based on paid alignment score across the 5,613 triangles**



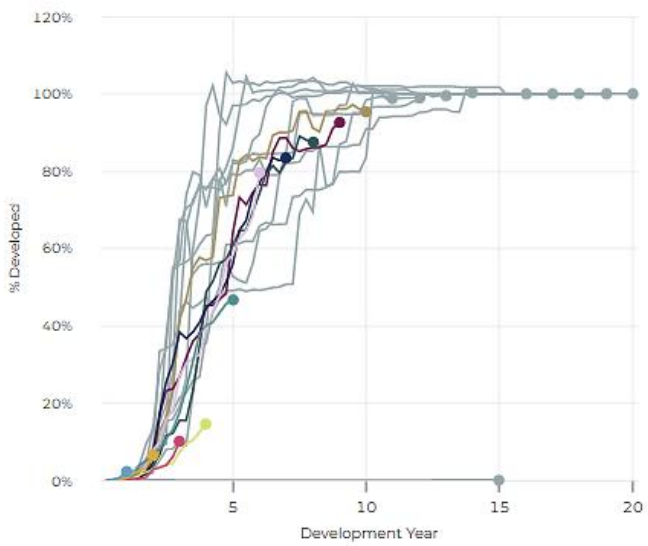

The results show a similar trend to the incurred data and clear evidence that paid alignment is correlated with observed reserve deteriorations. We see a weaker signal at the high alignment scores, with 43% of classes in the top bucket experiencing a reserve deterioration, compared to 68% on incurred data. We postulate two potential reasons for this:

- **Paid data is typically slower developing**, so it is possible that our average time horizon of 3 years is not allowing sufficient time for all deteriorations driven by paid claims to be recognised and emerge.
- **Incurred-based reserving methods are more common than paid based methods** across the general insurance market. This is because they capture both claim payments and outstandings. A diagnostic that is directly linked to the projection methodology would ordinarily be expected to be a stronger predictor of performance. Where ultimate claims are based primarily on an incurred method, the incurred to ultimate data is

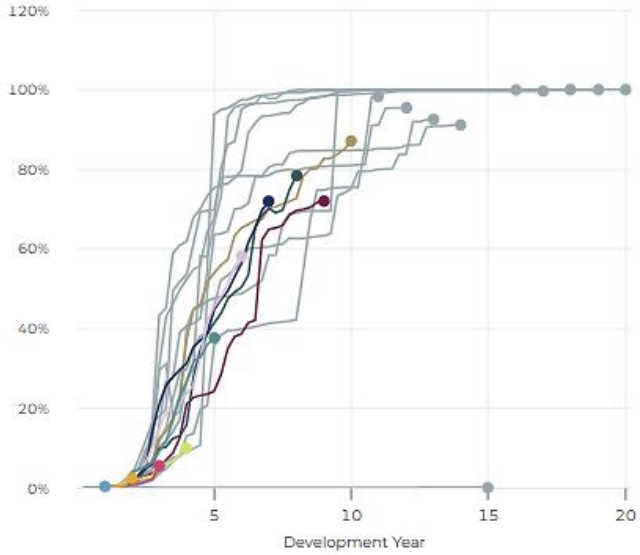
typically stable and the trends are easy to spot (figure 8 - below left). The paid data is often more scattered and volatile (figure 8 below – right) making it harder to identify trends against noise in the data.

**Figure 8: Comparison of incurred and paid to ultimate metrics where an incurred based projection is used**

Incurred as % of ultimate




Paid as % of ultimate




As with the incurred alignment score, we can use the results of our Schedule P analysis to calibrate paid alignment score for use on other datasets. Promisingly, the calibration is consistent with the incurred claims data, meaning a single scoring approach looks to work across both paid and incurred data – and suggesting that at least to some extent the scoring is metric agnostic.

**Table 3: interpretation of results from the paid alignment score**

| Paid alignment score | What you would expect to see in a diagnostic graph | Interpretation |
|---|---|---|
| $A \leq -6$ | Most cohorts sit well below past data | Likely prudent |
| $-6 \leq A < -2$ | Some early cohorts below past data and / or many cohorts towards bottom of past data. | Possibly prudent |
| $-2 \leq A < 2$ | Cohorts overall well aligned with past data | Approximately best estimate |
| $2 \leq A < 6$ | Some early cohorts above past data and / or many cohorts towards top past data. | Possibly optimistic |
| $6 \leq A$ | Most cohorts sit well above past data | Likely optimistic |

## 4.7. Extension to consider gradient alignment score

As an extension to the original testing, we observed circumstances where incurred to ultimate alignment looked reasonable and scored reasonably too, but where one or more accident periods were clearly developing on an upwards or downwards trajectory that looked likely to significantly worsen the score in the future. In other words, alignment score was reasonable at the evaluation date but looked highly likely to become less reasonable over future time periods.

To test if these situations are also predictive of reserve deteriorations, we define a metric called the gradient alignment score, using the following approach:

$$M^{\Delta}_{i,j} = \log\left(\frac{M_{i,j}}{M_{i,j-1}}\right) = \log\left(\frac{C^{\text{inc}}_{i,j}}{C^{\text{inc}}_{i,j-1}}\right) = \log Y^{\text{inc}}_{i,j-1}$$

Where $Y^{\text{inc}}_{i,j-1}$ is the link ratio between periods $j$ and $j-1$.

The reason we take logs in the above equation is because the incurred to ultimate gradient is a multiplicative relationship and therefore a doubling and a halving are treated as equal-and-opposite:

$$\log(2) = +0.693, \qquad \log(1/2) = -0.693.$$

The gradient alignment score can then be calculated by using the alignment function applied to the log link ratio:

$$A^{\Delta} = A(M^{\Delta})$$

The purpose of the gradient alignment score is to identify classes where the alignment score is trending upwards or downwards over recent time periods.

We performed the same testing analysis on the gradient alignment score, and the results are shown below.

**Figure 9: Incurred (left) and paid (right) gradient alignment**

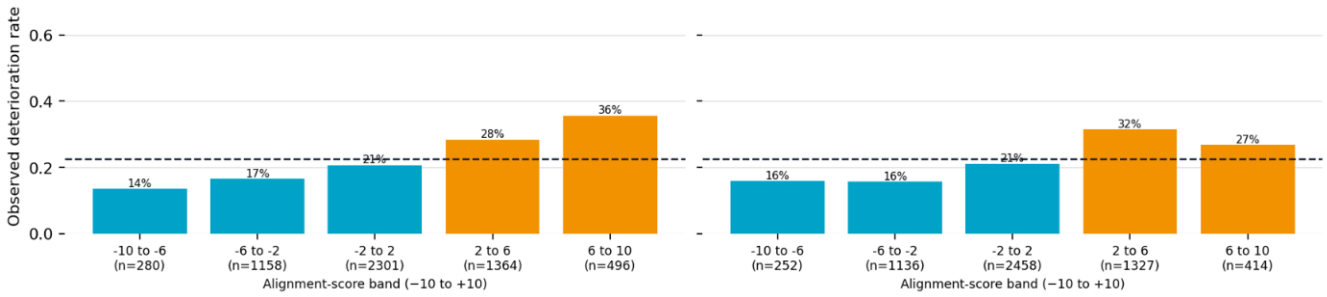



Figure 9 shows the performance of these gradient alignment metrics on the same underlying data. We can see that the incurred gradient alignment score exhibits a clear correlation with reserve deteriorations.

The results are more mixed for paid gradient alignment. As in section 4.6, we suspect this is primarily driven by paid to ultimate data being noisier because of the propensity of the market to use incurred-based techniques.

## 4.8. Discussion, conclusions and implications for actuarial work

Our analysis demonstrates that it's possible to compute a numerical incurred (or paid) alignment score that effectively represents the diagnostic information that an actuary would typically obtain via a graphical review of an incurred (or paid) to ultimate chart.

Both the incurred and paid alignment scores are shown to be effective predictors of reserve deterioration within the data that we tested. In addition, incurred gradient alignment was also shown to be an effective predictor, whereas paid alignment was of more marginal value.

The Schedule P data is broad multi-company and multi-class data for a large and mature insurance market. As such, results based on this data suggest that the approach may be applicable across a range of property and casualty reserving contexts, subject to the limitations discussed in section 4.9 and 10.

In practical terms, by using an alignment scoring framework, actuaries can reduce reliance on reviewing a large number of graphs when looking to review the reasonableness of reserving assumptions. Instead, the same information can now be expressed tabularly, with classes exceeding particular thresholds (eg those thresholds set out in tables 2 and 3) highlighted for review. We would expect this to save time when: performing summary reviews

of assumptions; when allocating IBNR to more granular classes for financial reporting; and also when rolling forward reserves month-to-month.

We also note that Incurred (or paid) alignment score as a stand-alone metric has some helpful advantages over AvsE analysis, which is currently the standard numerical framework that most firms use to assist with evaluating the reasonableness of their projections.

- Alignment score can be calculated for newly derived ultimate claims – you don't need to wait for experience to emerge (or conduct back-testing that has inherently circular logic). This means it can be used to evaluate newly derived ultimates, as well as on roll-forward.
- The standard deviation component of the alignment score takes into account volatility of past data – so larger variations between mean incurred to ultimate ratio and past data are permitted for more volatile datasets. When considering AvsE, these require investigation, eg where a large loss drives adverse AvsE, this may still be within normal volatility for a particular class.
- Incurred alignment score is method agnostic – whereas AvsE requires a development pattern even if this is not being used to set ultimate claims.
- The capping applied to each cohort (+/-5 standard deviations) limits the potential for a score to be biased by a single anomalous cohort of data.

## 4.9. Limitations

### Reliance on assumptions underpinning traditional actuarial techniques

It is important to note that although incurred (or paid) to ultimate position is considered by actuarial practitioners to be a good indicator of whether reserves are reasonable, it can be misleading in the following circumstances:

- Changes in initial case reserves (or payment patterns)
- Changes in initial claims handling methodology, particularly speeding up / slowdown of the process
- Changes in the effect of inflation over time
- Changes in mix of claims
- Changes in claims reporting pattern

In essence, like the actuarial techniques it depends on, this metric relies on the past being a good guide to the future, and on the dataset it is applied to being reasonably homogenous. This limitation is not unique to alignment scores but also affects the graphical review of the same diagnostics and other techniques such as AvsE.

### Assumption that incurred (and paid) to ultimate metrics are symmetrically distributed about the mean

We calculate the number of standard deviations from the mean in our alignment metrics and score them symmetrically. This implicitly assumes that deviations above and below the mean are equally likely and the distribution is symmetrical. In practice this may not be the case, particularly for large-loss exposed data. For example, there may be a large number of cohorts closely grouped below (or above) the mean and a small number of high (or low) outliers. The extent to which the distribution of incurred and paid to ultimate ratios is symmetric or not will depend on a range of class specific factors, such as potential for recoveries, large losses, catastrophe events, as well as the same company specifics discussed above. For skewed data, alignment scores may overstate the risk of deterioration in the direction of the skew, and understate the risk the other way.

Further overall limitations of our research are set out in section 10.

# 5. A multiple metric model

## 5.1. Overview

In this section we extend our analysis of metric alignment scores to build a machine-learning led framework that can predict future reserve deteriorations based on one or more such alignment metrics, together with any other KPIs that may be predictive of reserve deteriorations.

The advantages of embedding the alignment scores in a machine-learning led framework are that:

- For a single parameter model, it is possible to optimise the parameter threshold (eg the -10 to + 10 score) that predicts a certain level of reserve deterioration (eg, a more precise formulation of our calibration in tables 2 and 3).
- For a multi-parameter model, it should be possible to out-perform a single parameter model by drawing inferences from multiple different actuarial metrics.
- A multi-parameter model is extensible to capture further diagnostics that actuaries believe could be predictive of reserve deteriorations in the future.

In the remainder of this section we discuss: the data used for model training (5.2); the model training process (5.3), and the key results (5.4). Our validation process for our final model is set out in section 6.

## 5.2. Data

### Data source

Both the single-metric and multiple-metric models require training data containing claims development data (triangles) at the start of a period and the subsequent movement in booked ultimate claims over future time periods. As relatively few public datasets contain both claims triangles and ultimate claims, we used Schedule P data from the annual statutory returns filed by US property and casualty insurers who report into the “National Association of Insurance Commissioners” (NAIC).

The Schedule P returns provide annual paid & incurred claims triangles and ultimates by insurer, reserving class and accident year and cover a rolling ten-year history. For a three-year comparison, only seven accident years are present in both the opening and closing returns. We therefore excluded the three oldest cohorts from the opening return and the three newest cohorts from the closing return, ensuring that reserve movements are measured for a consistent set of accident years.

We used three of these cross-era datasets. These were chosen to ensure that our training data and evaluation data covered a range of different historical periods. The table below shows how we used the various data in each era. We describe an “era” as spanning the model fitting date (the start of the era) and finishing on the evaluation date where reserve deteriorations are assessed (the end of the era). This is summarised in table 4 below.

### Table 4: NAIC data for training and testing

| Schedule P data "era" | Prediction model fitted to triangles of the following data | Reserve deteriorations evolve (unseen by model) over | Model predictions tested as at year end: | Dataset usage | Observations (ie triangles available to train and test on) |
|---|---|---|---|---|---|
| 2008-2011 | 2002 - 2008 | 2009 - 2011 | 2011 | Model training | 3,040 |
| 2019-2022 | 2013 - 2019 | 2020 - 2022 | 2022 | Model training | 2,579* |
| 2022-2025 | 2016 - 2022 | 2023 - 2025 | 2025 | Model testing | 2,508* |

**see note below regarding management of overlap between 2019-2022 and 2022-2025 datasets.*

The schedule P data represents one of the largest publicly available datasets that includes both triangle and claims data for a large and mature insurance market, including a wide variety of companies and insurance business types. Training on a broad dataset should lead to a model that is broadly applicable, rather than being specific to a single narrow range of business groupings.

We note that there is overlap between the cohorts in the 2019-2022 training and 2022-2025 testing data. To reduce the risk of overlap between these two datasets, we split each of these datasets into two groups, with different combinations of companies and reserving classes in each. We then used one set of classes in the 2019-2022 dataset for training (discarding the other group) and then tested on the opposite grouping. This is discussed in further detail in the training, testing and validation splits section.

## Data cleaning

We applied the same data-cleaning rules to each of the three datasets described above. The purpose here was to remove:

- small classes exposed to higher claims volatility;
- classes whose diagnostics could not be reliably trusted;
- duplicate entries due to reporting requirements; and
- extreme deteriorations that were unlikely to be caused by under-reserving and more likely caused by data corrections.

To achieve these objectives, we applied the following data cleaning rules:

- We restricted the analysis to material entities with at least $10 million of opening reserves and at least six cohorts with non-zero ultimate claims.
- We removed entities where there were duplicate incurred to ultimate development triangles. These generally represented quota-share or group-company copies.
- We removed entities where material movements year-on-year looked to be due to data corrections rather than actual year-on-year movements.
- Where premium was not reported, premium growth was assigned a value of one, representing no change.

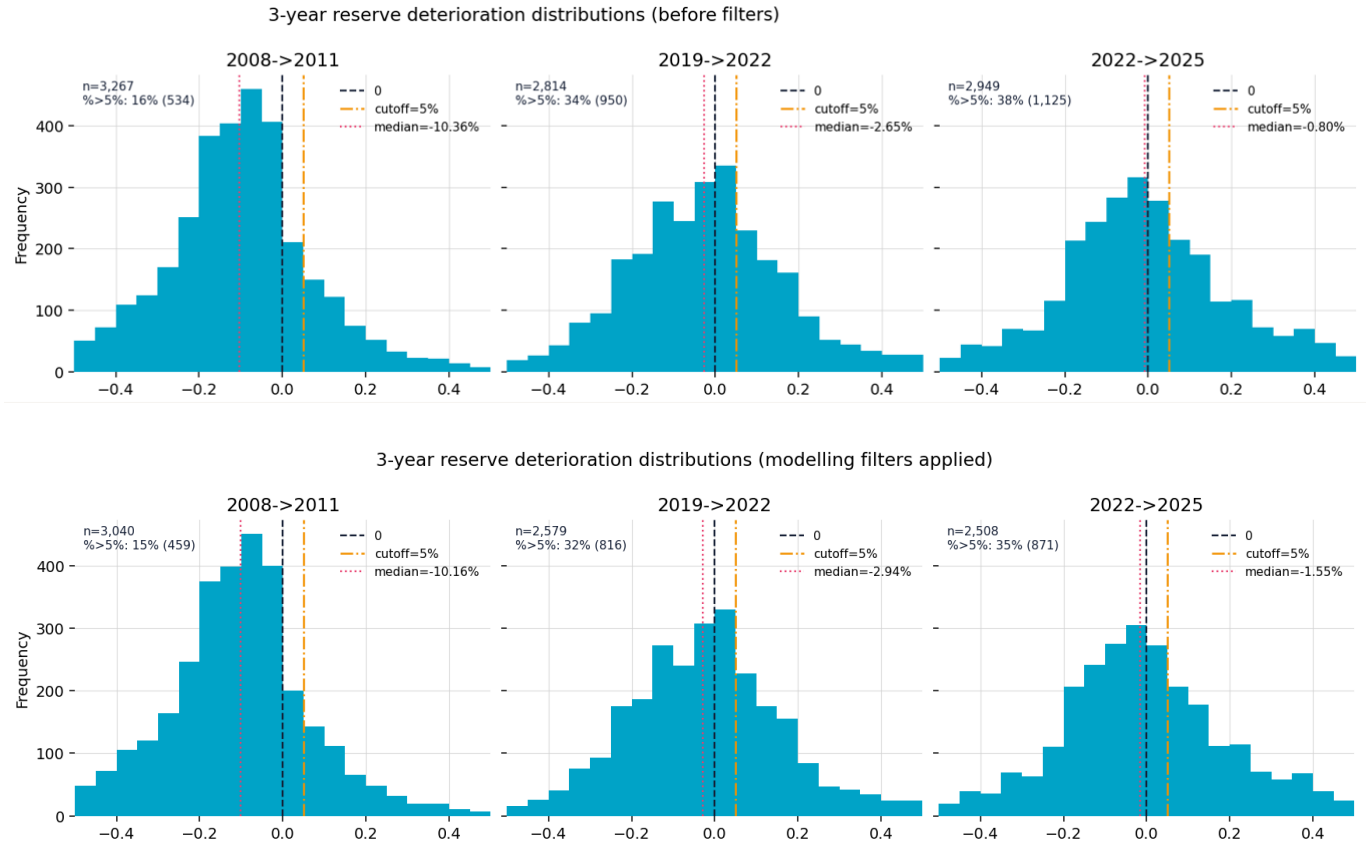
Following these checks, the three eras contained 3,040, 2,579 and 2,508 usable observations respectively.

We then reviewed the complete distribution of 3-year reserve movements before and after filtering.

- The proportion of entities deteriorating by more than 5% was approximately 15%, 32% and 35% respectively across the three eras. These differences reinforce the importance of testing the model on a mixture of calendar periods and of assessing probability accuracy relative to each dataset's underlying deterioration rate, which we discuss in the model-assessment section below.
- Figure 10 shows the distribution of reserve movements. It is desirable to use data that:
  - has broadly equal numbers of reserve improvements and deteriorations (ie is broadly symmetrical)
  - has median reserve movement of close to zero – ie suggesting that the reserves set in the underlying data are approximately best estimate, and not materially prudent or optimistic.

We note that the 2008-2011 era has a median reserve movement of -10.16% after filtering and a positive skew. Validation testing in section 6.2 (table 11) demonstrates that this does not appear to have an adverse effect on model training.

**Figure 10 – before and after distribution of reserve deteriorations within the 3 Schedule P returns used for training, testing and validation**



### Training, testing and validation splits

The 2008-2011 and 2019-2022 eras from table 4 were used for model training with the 2022-2025 era reserved for out-of-sample testing, since it's clearly desirable to test the model on different data to that used for training.

#### *Avoiding leakage between training and test data*

To avoid leakage between the training and testing data (ie overlap between the datasets), we identified any reserving classes (combinations of company and business group) that were present in both the training and the test data and removed them from one or the other. Specifically

- 60% of these reserving classes were retained in the training data and discarded from the test data.
- 40% were allocated to testing data and discarded from the training data.

#### *Initial parameter training data*

When fitting the model to the two training data eras, 80% of the training data pool (after the above adjustments) was used to set the underlying model parameters.

The model was therefore fitted to 2,533 reserving classes, which included 3,514 observations in total, because some reserving classes are present in both of the two training data eras.

#### *Operating threshold setting data*

A 20% entity-grouped validation sample was held back from the training data. This was used to select the $F2$ operating threshold (the cutoff score above which a class was considered a reserve deterioration risk). The setting of the operating threshold is discussed in more detail in the section on model assessment.

#### *Testing data*

The selected model and operating threshold were then applied to the fixed 2022-2025 out-of-era test set of 1,314 observations to examine performance on data it had not seen previously.

All candidate models were evaluated using exactly the same validation and test observations, where each feature could be calculated.

## 5.3. Modelling

### Model choice and design

We framed model training as a binary classification problem.

Each observation represented triangles and ultimates for one company and reserving class combination.

Using only information available at the start of the period, the model predicted whether booked ultimate claims would deteriorate by more than 5% of opening reserves over the following three years.

Eg for the 2008-2011 dataset, the model would:

- Make a prediction of whether the company / class would deteriorate by 5% or more, based on triangulated data covering 2002-2008.
- That prediction would then be tested against the data as at 2011 (ie 3 years later) to see if a 5% or more deterioration had actually occurred.
- Data between 2009 - 2011 was used only to determine the outcome and was never included within the modelling.

We tested three candidate model types:

- Random forest
- Logistic regression
- Logistic regression with interaction terms

Each candidate model was tested using consistent data and features. The selected specification was an unpenalised logistic regression containing five main effects:

- Incurred-to-ultimate alignment
- Paid-to-ultimate alignment
- Incurred-to-ultimate gradient alignment
- Paid-to-ultimate gradient alignment
- Premium growth over five years

Interaction terms were investigated during development, but after reviewing the 2D ALE plots and regression diagnostics, the final fitted model did not include any interaction terms.

The simple logistic regression was the resulting preferred model since it achieved comparable out-of-sample performance whilst remaining more transparent and easier to interrogate than the random-forest alternative.

The final multi parameter model is set out below:

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -1.50 + 0.85X_{1,i} + 0.27X_{2,i} + 0.17X_{3,i} + 0.11X_{4,i} + 0.20X_{5,i}.$$

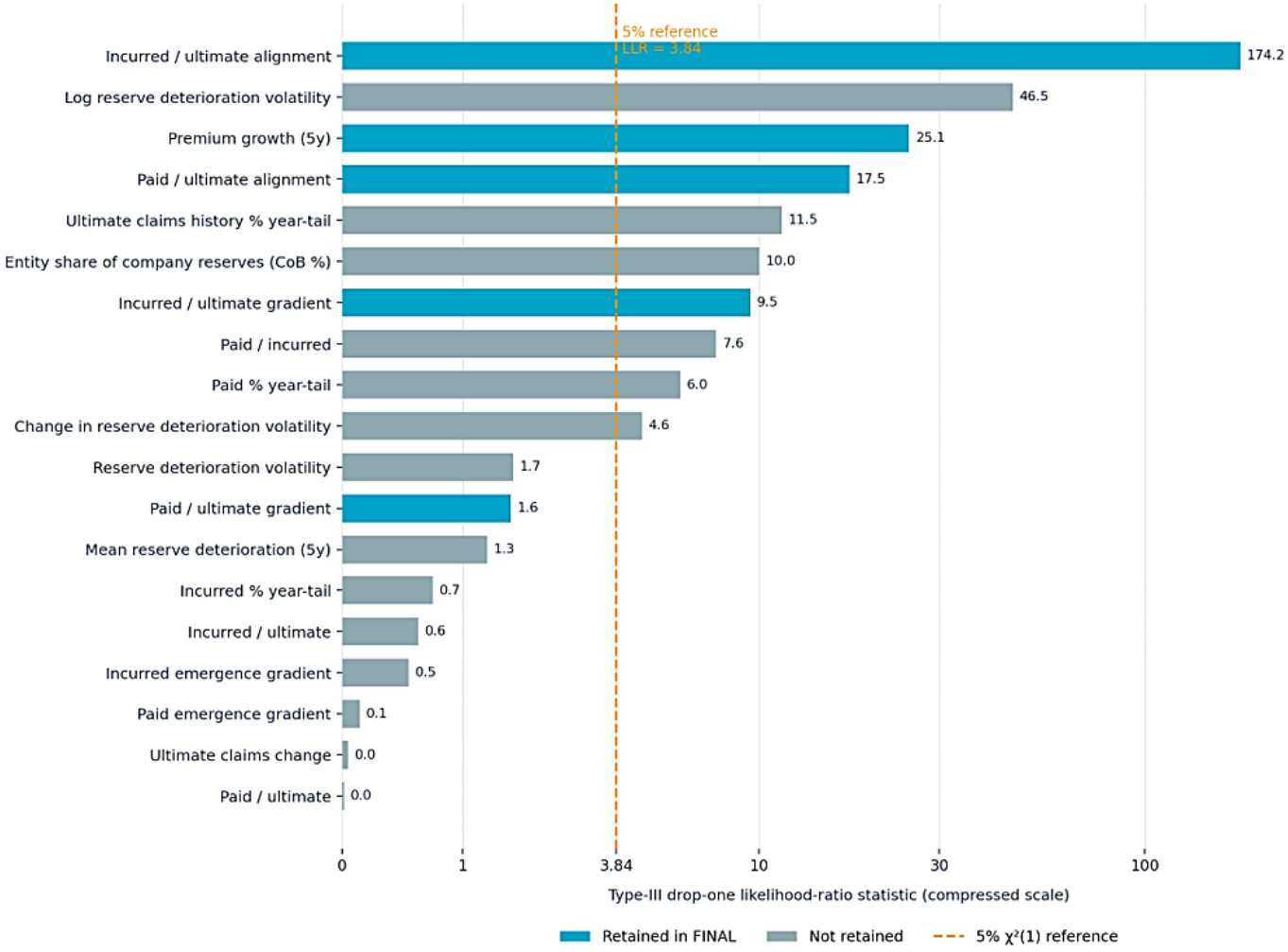
$$\begin{aligned}
\hat{\boldsymbol{p}}_{\boldsymbol{i}} &= \text{Reserve deterioration risk score},\\
\boldsymbol{X}_{\boldsymbol{1,i}} &= \text{Incurred / ultimate alignment},\\
\boldsymbol{X}_{\boldsymbol{2,i}} &= \text{Incurred / ultimate gradient alignment},\\
\boldsymbol{X}_{\boldsymbol{3,i}} &= \text{Paid / ultimate alignment},\\
\boldsymbol{X}_{\boldsymbol{4,i}} &= \text{Paid / ultimate gradient alignment},\\
\boldsymbol{X}_{\boldsymbol{5,i}} &= \text{Premium growth (5y)}.
\end{aligned}$$

The coefficients were calculated using Newton-Raphson maximum likelihood estimation.

## Feature selection

We started with a broad set of candidate features covering a range of standard reserving diagnostics, cohort-alignment measures and static reserving information that could be obtained or derived from the Schedule P returns. Figure 11 below illustrates feature importance within a preliminary model

**Figure 11: feature importance of candidate model features** *(note compressed scale)*




Forward-era testing showed that the incurred and paid alignment metrics contained most of the predictive signal from the typical graphical reserving diagnostics, and adding the remaining features and their interactions did not statistically improve performance. We therefore focused on the alignment metric and individual-metric feature families.

We also removed highly correlated features (eg the incurred and paid emergence scores) and a number of more highly predictive metrics were excluded for practical reasons as described in table 5 below to ensure our framework could be applied more broadly to other actuarial datasets. Correlation is discussed later in this section.

The paid / ultimate gradient alignment metric was included for reasons of symmetry (ie because the equivalent incurred / ultimate metric was included), despite being shown to be of limited predictive power in the test case evaluated. We consider this important as we expect this metric to be more predictive for classes where reserving is based on a paid method rather than an incurred method.

### Table 5: justification for excluding highly performing metrics

| Reason for excluding metrics | Metrics excluded | Discussion |
|---|---|---|
| **Ease of calculation and availability of data** | • log reserve deterioration volatility<br>• ultimate claims history as % of tail<br>• Change in reserve deterioration volatility | They are easily available (though more complex to calculate) within the NAIC data, but not metrics that we expect actuaries to automatically have easy access to |
| **Ensuring reserving classes can be evaluated independently** | Class reserves as proportion of company reserves | Including this metric introduces a dependency between the modelling of a specific class and the overall reserving exercise, which is undesirable. Our view is that it is preferable to evaluate each class on a standalone basis. |

## Model assessment

We assessed candidate models (across model types and feature subsets) using three complementary measures:

- **AUC** [6] measures how effectively the model predicts events (deteriorations) compared to non-events (no deterioration) without relying on a defined probability threshold. This can be framed as the probability that a randomly selected deterioration is predicted (to deteriorate) with higher probability than a randomly selected non-deterioration – ie the model is working well and directionally as expected:

$$\text{AUC} = \Pr(\hat{p}^+ > \hat{p}^-) + \frac{1}{2}\Pr(\hat{p}^+ = \hat{p}^-)$$

  Here, $\hat{p}^+$ is the predicted probability for a deterioration and $\hat{p}^-$ for a non-deterioration.

- **Brier scores** [7] measure the mean squared difference between the predicted probability and the observed binary outcome. A lower Brier score indicates that the projected probabilities (in this case that a class will experience a reserve deterioration) are close to the observed binary outcomes (ie whether a class does or does not deteriorate).

  When the prevalence of a deterioration differs between time-periods, we compare the score with a simple "base-rate" forecast instead, using the Brier skill:

$$\text{Brier}_{model} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$$

$$\text{Brier}_{\text{base}} = \frac{1}{N}\sum_{i=1}^{N}(\bar{y} - y_i)^2 = \bar{y}(1-\bar{y})$$

$$\text{Brier skill} = 100\left(1 - \frac{\text{Brier}_{\text{model}}}{\text{Brier}_{\text{base}}}\right)$$

  where $y_i \in \{0,1\}$ is the observed outcome and $\hat{p}_i$ is its predicted probability. The base rate forecast is a simple "no-skill" benchmark to allow for differences in the number of deteriorations that did occur in the dataset being analysed. If this isn't acknowledged, an overly confident model sampled on a largely adverse dataset would show as misleadingly predictive relative to other models tested on other samples.

- **Recall** measures the proportion of deteriorations correctly predicted (count of true positives, "TP") as a proportion of the total reserve deteriorations (count of true positives + false negatives, "TP + FN").

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Precision** is a counterpart defined as $\frac{\text{TP}}{\text{TP} + \text{FP}}$

FP is the count of false positives, these are the cases that the model flags as a deterioration risk but where no deterioration is actually observed over the three year time horizon.

Considering all three measures is important. A model can achieve a strong AUC by ranking entities correctly while still producing inaccurate probabilities or performing well at some operating thresholds (points on the ROC curve where a model score $\geq$ this value is classified as a deterioration) whilst not at those that are relevant for the forecast in hand. Model selection is therefore not based on AUC alone.

## Operating threshold

The output of the logistic regression model is a percentage confidence that any particular class will experience a deterioration. For a binary classification problem, we need the model to determine whether the class will (or will not) deteriorate, based on that probability. We do this by assigning an operating threshold, which is a score above which a class is predicted to deteriorate and below which it isn't.

The operating threshold does not affect the fit of the model, changing it primarily adjusts the mix of false positives to false negatives.

- Setting a low operating threshold means classes with more marginal characteristics will be flagged for review, likely increasing the number of classes flagged that do not subsequently deteriorate (false positives).
- Setting a high operating threshold means only very clear cases of deterioration risk will be flagged, but as a consequence a higher proportion of moderate examples will be missed (false negatives).

We selected the operating threshold $\tau^*$ using the $F\beta$ [8] score with beta equal to two, commonly referred to as the $\text{F2}$ score, where:

$$F\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

The threshold $\tau^* = \underset{\tau}{\text{argmax}}(F\beta(\tau))$ was set using the model trained on 80% of the training data and evaluated on the remaining 20% hold-out sample.

The threshold that maximises $\text{F2}$ was 14% deterioration risk score.

Unlike $\text{F1}$, which gives equal importance to precision and recall, $\text{F2}$ places greater weight on recall. This reflects the intended use of the model - missing a genuine deterioration is considered more costly than asking an actuary to review an additional false-positive case.

The resulting confusion matrix (see table 8) is therefore asymmetric by design. We deliberately accept more false positives where this reduces false negatives.

Changing from $\text{F1}$ to $\text{F2}$ does not change the fitted model or its AUC; it changes the point at which an entity is flagged for review to adjust the number of false positives relative to false negatives. In practical terms, optimising to $\text{F1}$ will reduce actuarial workload by flagging fewer classes. As a result, some predictable deteriorations will be missed. Optimising to $\text{F2}$ will flag more classes, increasing workload, but reducing the risk of missing classes at risk of deterioration.

## 5.4. Results

Table 6 below shows the results of our final fitted model applied to the 1,314 reserving classes in the 2022-2025 testing dataset. It is also compared to the performance of fitting the model to just the incurred alignment score metric – the strongest metric considered in section 4.

### Table 6: Model prediction results

| Model | Actual deteriorations observed | Deteriorations predicted correctly | % success *(recall)* |
|---|---|---|---|
| Logistic regression Incurred alignment only | 476 | 370 | 78% |
| Logistic regression multi-parameter model | 476 | 388 | 82% |

At first glance, the incurred alignment only model predicted 78% of deteriorations correctly and the multi-parameter model performed slightly better at 82%, both of which are promising results. Of course it is possible to predict 100% of deteriorations simply by flagging 100% of classes as at risk of deteriorating, but this does not address the ultimate goal of reducing actuarial review time!

An equally valuable consideration therefore is the reduction in number of classes flagged for review by the model, ie those classes that are perceived to be at risk of deterioration, whether or not they subsequently experience a deterioration. Table 7 below extends the analysis to consider the reduction in workload for a reviewing actuary by focusing the review on classes flagged as a deterioration risk by the model. Note this is based on setting the operating threshold by maximising $F2$ so we inherently accept a relatively high number of classes to review as the trade off for predicting a very high proportion of deteriorations.

### Table 7: prediction results including actuarial review time improvement

| Model | Actual Deteriorations Observed | Deteriorations predicted correctly | % success *(recall)* | No. classes flagged for review | Total classes | % classes flagged for review |
|---|---|---|---|---|---|---|
| Incurred alignment only | 476 | 370 | 78% | 795 | 1,314 | 61% |
| Multi-parameter model | 476 | 388 | 82% | 824 | 1,314 | 63% |

*Based on this analysis we can see that, when using our multi-parameter model, we can:*

*a) Flag 82% of actual reserve deteriorations of 5% or more correctly, and*

*b) Reduce the number of classes requiring review by 37%.*

Without accounting for review of the numerical diagnostics themselves being more efficient than a manual graphical review process (see section 7 for more details on why this might be the case), a 37% reduction in classes to review already translates into a significant improvement in actuarial review vs a "review everything" approach.

We note the 37% reduction in classes to review assumes that in the absence of using this approach, the incurred and paid to ultimate charts would be evaluated graphically instead for every class. In practice, the time saving may be less than 37% because:

- Some of the classes not flagged by our model would only require a very brief check by an actuary. Most of the time is spent on more complex cases that would likely be flagged as a deterioration risk.
- It may be possible to confidently rule out some unflagged classes using other simple diagnostics present in existing review frameworks.

We note that the inherent volatility in insurance business means that we will never be able to predict 100% of future reserve deteriorations based on past data, eg due to latent claims, legal judgements, large losses and catastrophe events. As such, we should not evaluate the 82% of classes predicted correctly against an upper bound of 100%.

Similarly, it is not possible to reduce the number of classes to flag for review to 0%. Even a perfect model would flag 36% of the test data for review, because that is the proportion of classes that actually went on to deteriorate by 5% or more and so require review and adjustment . However, there is scope to improve the precision of the model, as long as this is possible without significantly sacrificing recall.

### Breakdown of prediction results

An alternative way of representing the trade-off between flagging more classes for review and detecting more classes at risk of deterioration is using a confusion matrix (see table 8). This compares actual results vs predictions in a quadrant analysis, looking at both true positives / negatives, but also false positives / negatives.

### Table 8: Example confusion matrix

| | Not Flagged | Flagged |
|---|---|---|
| No Actual Deterioration | True Negative (TN) | False Positive (FP) |
| Actual Deterioration | False Negative (FN) | True Positive (TP) |

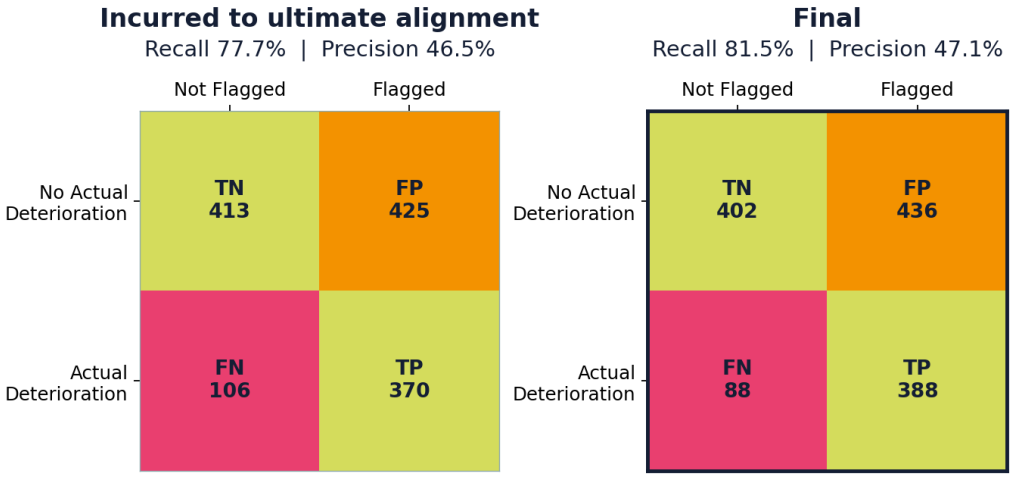
An ideal model maximises detection of true positives and negatives, whilst avoiding false results. As previously discussed, there are two ways to achieve this, and the optimal approach depends on the purpose of the model:

- **F1 maximisation** looks to minimise the total number of false flags, irrespective of whether they are false positive or negative
- **F2 maximisation** looks to minimise the total number of false negatives but accepts more false positives (and often total false results) as a consequence.

From an actuarial perspective, this algorithm will be used to flag classes at risk of deterioration for a more detailed human review. Therefore, it is likely better to minimise false negatives, rather than total false results. This is because the false negatives are at-risk classes that will not be flagged for review. On the other hand, false positives will be flagged for actuarial review and (perhaps quickly) confirmed as reasonable by the human in the loop.

For this reason, in our analysis we optimised the algorithm to maximise F2. Tables 9 and 10 below show the confusion matrices for our two models following F2 maximisation. The selected model is shown in the table and charts below with a bold outline.

### Table 9 & 10: Confusion matrices for our final models




We note that after F2 maximisation, there remains an 18% false negative rate (false negatives / total deteriorations) on our multi parameter model. We believe this primarily reflects the fact that insurance business is inherently uncertain and not every reserve deterioration can be predicted based on the data.

### Model performance evaluation

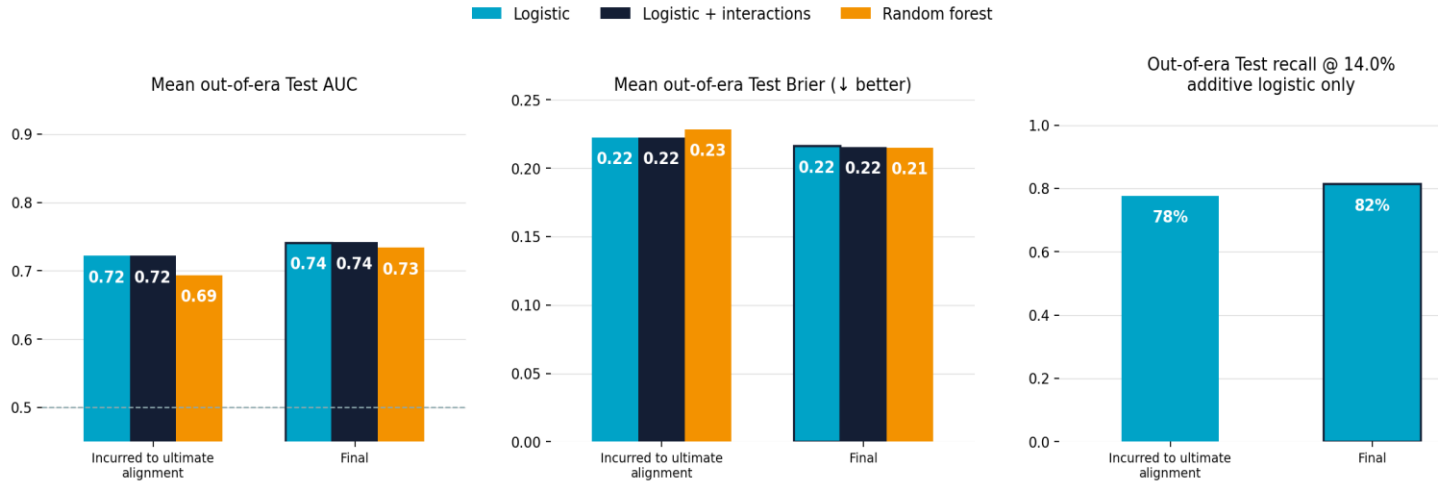
Figure 12 below shows the results of our model testing, evaluated using a combination of AUC, Brier score and Recall. These results are based on the average model performance across all our training and testing runs.

We performed testing for three different candidate models against each metric. We applied each type of model to both:

- Our final multi-parameter model
- Incurred alignment score only – as the most predictive single metric

The rationale for evaluating incurred alignment on a stand-alone basis is that section 4 illustrated that this was the strongest individual predictor of reserve deteriorations that we investigated. Actuaries looking for a simple framework to implement may benefit from the analysis of incurred alignment on a stand-alone basis. More mature teams may benefit from implementing (and extending) the full model.

## Figure 12 – AUC, Brier and recall results by model type and features




We note both parameter sets performed well in our modelling across all model types. Both types of logistic regression performed similarly and slightly better than the random forest model, though not by a significant amount. We selected the simple logistic regression model as our preferred model, given the greater transparency of the parameter weights, compared to using a random forest.

An interesting observation is that the 5-parameter model only marginally outperformed the single parameter incurred to ultimate model – which is at first glance a surprising conclusion. We performed further analysis to investigate the relative importance of features in the multi-parameter model and understand this result in more detail.

### Feature importance

We assessed the overall importance of each feature using Type III likelihood-ratio (LLR) statistics. We first fitted the complete model and then removed each feature in turn, refitting the model while leaving all other features unchanged.

- If the reduced model fitted the observed outcomes almost as well, the omitted feature was providing little information beyond that already contained in the other variables.
- Conversely, a large deterioration in model fit produced a high LLR statistic, indicating that the feature contributed important and non-duplicative information.

Because every feature is tested against the same complete model, the results are independent of the order in which features are considered. Figure 13 shows the relative importance of all 5 features of our multi-parameter logistic regression.

### Figure 13: Feature importance




The feature importance analysis confirms that the incurred to ultimate alignment metric is significantly the most influential predictor within this fitted model . The fact that it is the most predictive metric is not surprising given the testing on this metric on a stand-alone basis in section 4.

That said, it is surprising the extent to which incurred / ultimate alignment dominates other terms, in particular the paid metrics, which in section 4 were shown to be reasonably predictive. We have identified two possible explanations for this:

- **Bias towards incurred methods when performing actuarial work:** As discussed in section 4, we postulate that this is due to a bias towards incurred techniques being used more widely in actuarial practice, making the incurred / ultimate alignment less volatile (see section 4, figure 8) and also more directly coupled with the reserving outcome. This would carry through to this being a more important feature in the multi-parameter model.
- **Correlation between incurred and paid metrics:** This is expected, since incurred = paid + outstanding. By definition it contains all the information in the paid data. Correlation is analysed more in the next section.

Despite its lower predictive power when assessed against the market-level Schedule P dataset, we would recommend including paid alignment scores in a general model, as there will be situations where:

- Reserving is based on paid metrics (eg due to recent changes in case estimation philosophy), in which case the incurred vs paid feature importance may be reversed.
- Incurred or outstanding claims data isn't available.

As such, a model that includes both paid and incurred metrics should be more generalisable than one that relies on incurred metrics only.

## Correlation between features

An ideal model includes very little correlation between different features. This is because avoiding strong correlation between features can improve coefficient stability and make individual feature effects easier to interpret.

We analysed correlation using a Spearman rank-correlation analysis, which can help indicate whether two variables tend to move together in a consistent way.

Spearman correlation is a rank correlation approach, which considers the ordering of the variables, rather than their actual numerical values, and so is preferred for measuring correlations between different types of variables.

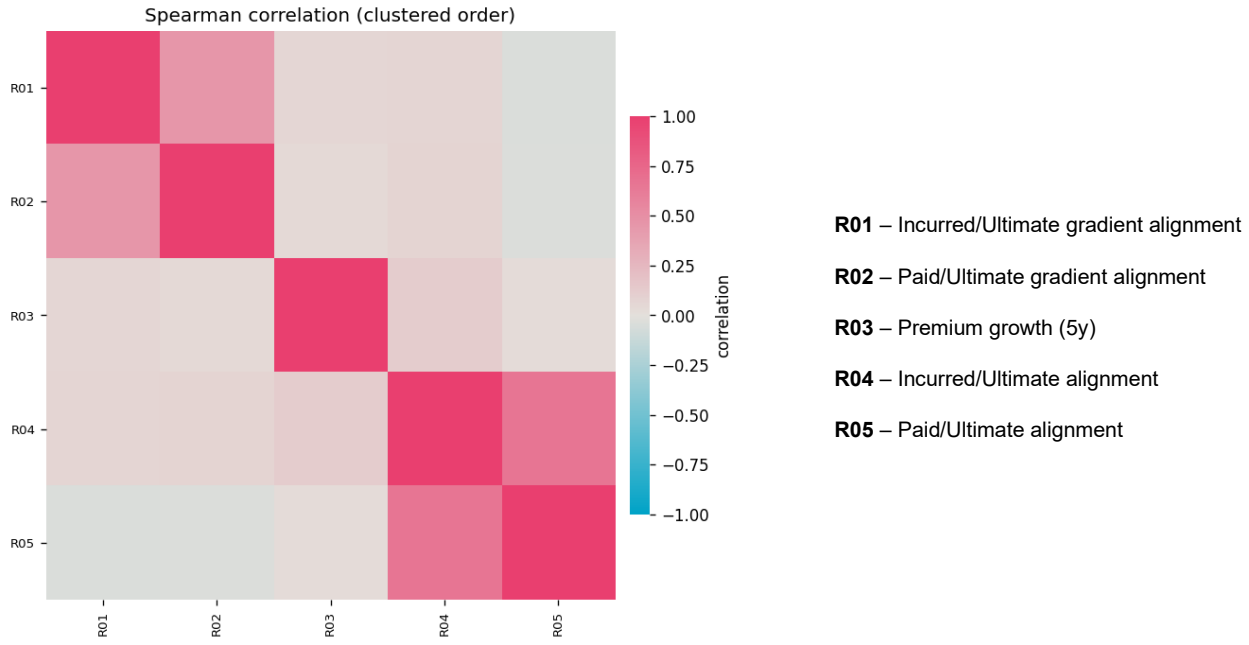
The results of our analysis are displayed as a heat map in figure 14 below. We can see that broadly there is little correlation between variables, except the following pairs:

- "Incurred to ultimate alignment" and "paid to ultimate alignment"
- "Incurred to ultimate gradient" and "paid to ultimate gradient"

It is expected that there will be some degree of correlation between these pairs. This is because incurred claims data consists of the paid claims + outstandings. Therefore, there will frequently be situations where the incurred and paid diagnostics (and hence alignment scores) behave similarly. The same applies to gradient alignment.

Despite the correlation between paid / incurred metrics, our view is that it is desirable to keep them both in the model, rather than relying on the more predictive incurred metrics. This is because (as previously discussed) we suspect that there is a bias towards using incurred methods to set ultimate claims, but where ultimates are based on paid data, the predictive power would likely reverse.

**Figure 14: Spearman correlation matrix for key model features**




## 6. Model validation

When training a model, it is important to confirm that its performance is generalisable or at least replicable under a broad range of conditions, rather than specific to the training dataset or initial configuration and parameters. To demonstrate this, we considered 5 validation questions:

1. Do we have enough data to train the model?
2. Does model performance depend on the specific data used for training and testing?
3. Does the model perform well across all business classes, including both short and long tailed lines?
4. Does the reserve deterioration threshold influence model performance?
5. How do minor changes to the alignment score formula affect model performance?

For each question we use an objective test procedure to evaluate the model, as set out in the subsections below.

## 6.1. Do we have enough data to train the model?

The selected model was fitted on progressively larger samples and evaluated both on those training observations and on the fixed 2022–2025 test set.

We assessed data sufficiency using entity-grouped learning curves. This means where an entity appears in two different eras, both entity entries are included in the same sub-sample. Performance from more data isn't then unduly influenced by including repeated entities from different eras.

Performance was measured using Brier skill: the improvement in probability accuracy relative to assigning every entity the same base-rate probability – see section 5.3 for a formal definition. We also showed results using the false-negative rate as an alternative error rate. The results are shown in figure 15 below.

### Figure 15 - Learning curve analysis

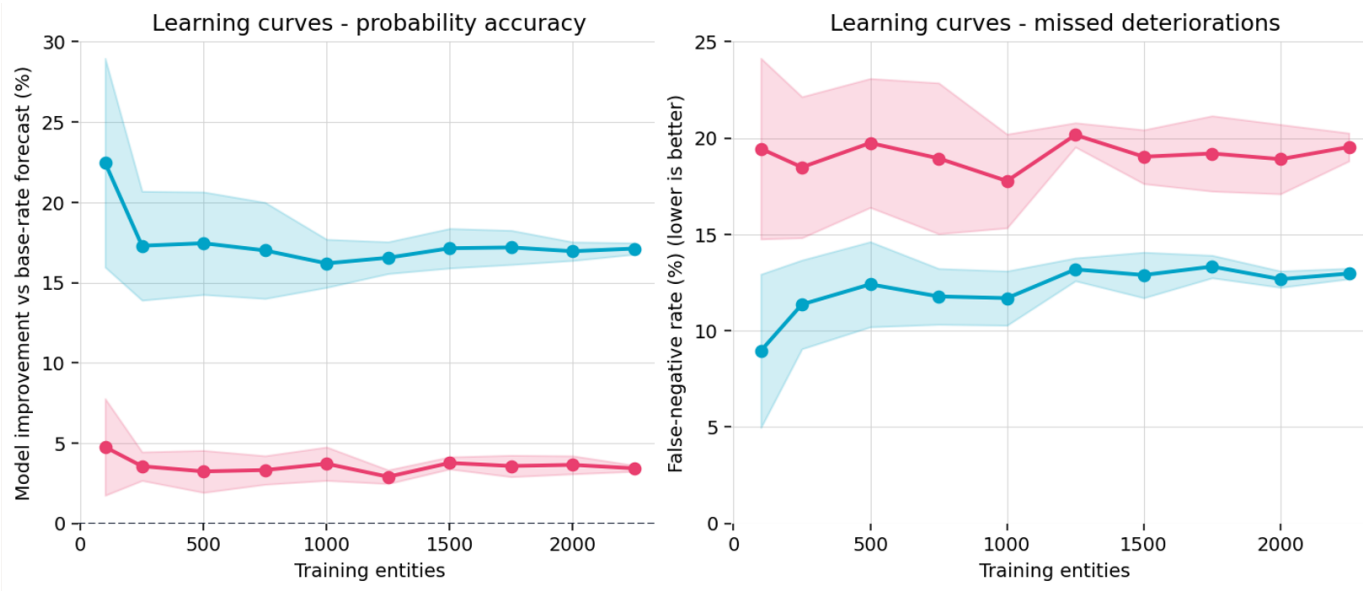



Figure 15 shows that model performance was initially volatile for small data samples but stabilised once approximately 1,000 training observations were included, with little further improvement as the sample increased towards 2,000 observations or beyond.

This shows that our training dataset of 3,514 observations is comfortably sufficient for our training purposes. Note that whilst this analysis shows further Schedule P data from our two eras is unlikely to improve model performance, other Schedule-P eras or other datasets may still improve performance if included in the training data..

The apparent need for 1000+ classes to stabilise the model is an important observation, because it suggests that, if the learning curve behaves similarly for company data, then larger data samples are needed for model training than many companies are likely to have internally.

The Brier learning curve (left) also shows a noticeable generalisation gap between the two curves (blue = training data, red = testing data) which reflects both ordinary entity-level generalisation and changes in the reserving environment over time. In particular any thematic class-level deteriorations will show better performance in the training data (since era-level class performance can be inferred from entities in that set).

The larger gap in Brier skill suggests that predicted deterioration probabilities are less stable across eras than the model's post-threshold binary classifications (ie deterioration or no deterioration predicted). The latter is closer to how we propose the model should be used in practice.

## 6.2. Does model performance depend on the specific data used, or is it more generalisable?

We assessed how generalisable model performance was through training and testing the model on different combinations of the underlying data and comparing model performance. The table below shows the different combinations of training and testing data considered, together with the AUC. The pink dataset is the testing run for which results are reported in section 5.4.

Overall, we considered 8 different combinations of re-training and testing including both forward and backward tests. Our chosen model demonstrates consistently high AUC across all tests, although there is some variation between the results of each test. This suggests that the approach is generalisable and not specific to any particular training or test data.

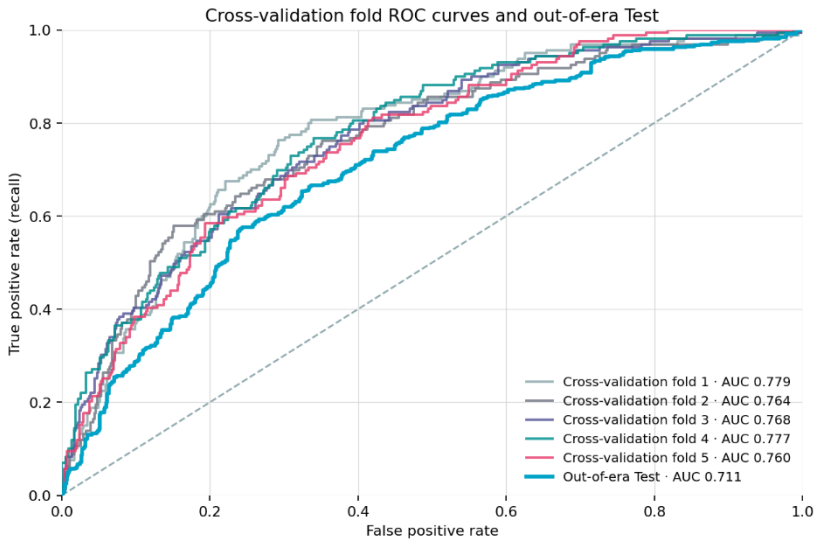
We note that all data ultimately comes from the Schedule P return. It would be valuable to test the model on market or company data from areas outside the US – noting that to some extent reserving practices vary geographically.

**Table 11: multi-parameter model performance when training and testing data is varied**

| Train | Test | Direction | Test AUC |
|---|---|---|---|
| 2008_2011 | 2019_2022 | Forward | 0.788 |
| 2019_2022 | 2022_2025 | Forward | 0.736 |
| 2008_2011 | 2022_2025 | Forward | 0.731 |
| 2008_2011+2019_2022 | 2022_2025 | Forward | 0.711 |
| 2022_2025 | 2019_2022 | Backward | 0.784 |
| 2019_2022 | 2008_2011 | Backward | 0.726 |
| 2019_2022+2022_2025 | 2008_2011 | Backward | 0.722 |
| 2022_2025 | 2008_2011 | Backward | 0.719 |

We see a similar story when we perform 5-fold out-of-entity cross validation on the training data – see figure 16, and that the AUC is not unduly affected by the inclusion or exclusion of individual entities in the training and test data subsets.

**Figure 16: ROC curve across 5-fold cross validation**



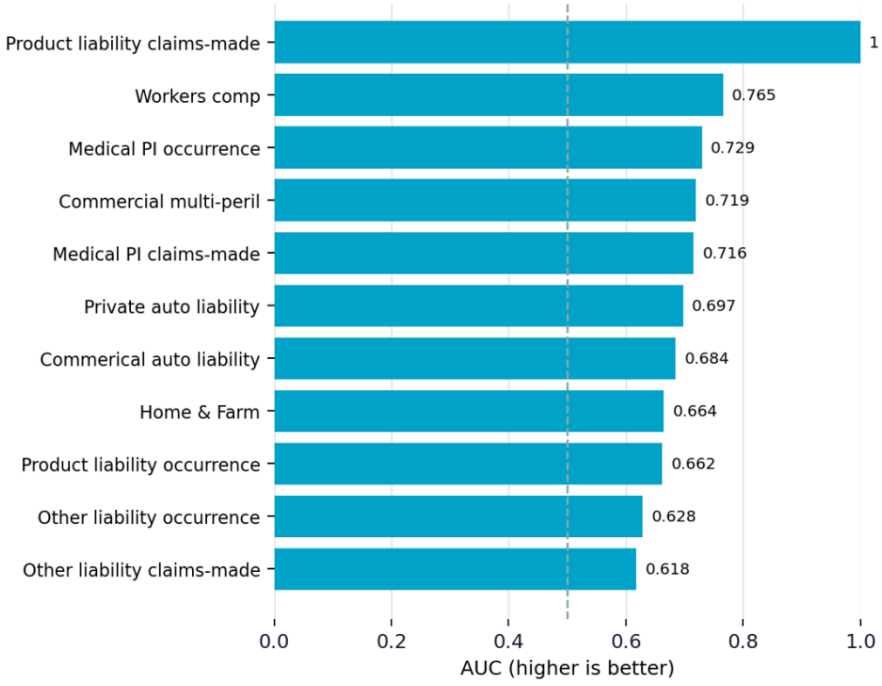
## 6.3. Does the model perform well across all business classes, including both short and long tailed lines?

A key question is whether our model can predict reserve deteriorations consistently across a wide range of classes, or whether it predicts well on some but loses predictive power altogether on others.

Figure 17 below shows the AUC by reporting class within the Schedule P data. Whilst there is some variation between classes, we see consistent performance above the threshold AUC of 0.6, demonstrating that the model has at least moderate performance across a range of classes in a mature and established insurance market.

We note an AUC of 1.000 on product liability claims made. There were only 17 test entities in this class, so this is likely to be an outlier, rather than an example of abnormally good model performance.

**Figure 17 – model performance by NAIC class of business**



## 6.4. Does the reserve deterioration threshold influence model performance?

Figure 18 below shows how the AUC of our model varies as a function of reserve deterioration threshold (ie how big a reserve movement is classified as a deterioration for our analysis). Note this is not the same thing as the model operating threshold. We selected a threshold of 5% for our analysis as a practical threshold above which actuaries and business were likely to be concerned.

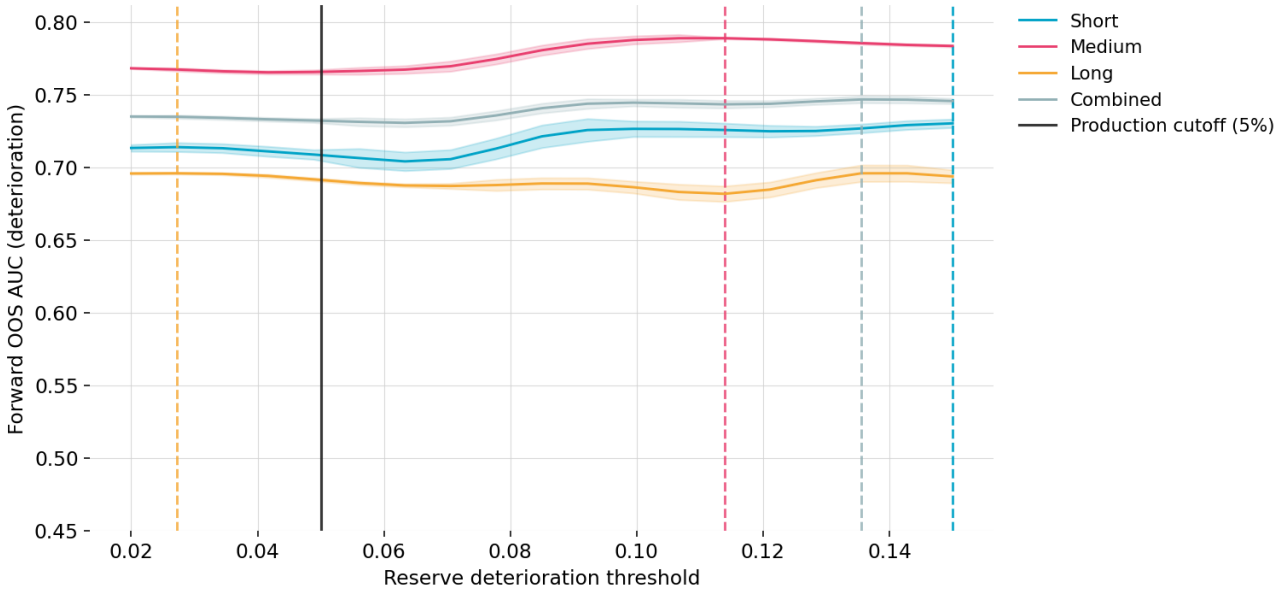
Our analysis shows that model performance is broadly consistent across different reserve deterioration thresholds – ie practitioners could use the model with a threshold aligned to company-level risk tolerances, rather than being limited to 5%.

Note that the combined AUC at 5% in the chart below differs from the mean AUC reported in Figure 12. This is because figure 12 represents the average AUC across multiple model runs, whereas this analysis was performed for a single training / test data combination.

We also considered how model performance and deterioration threshold varied by the nature of the business – specifically whether the business was long-tailed, medium-tailed or short-tailed. This is an extension of class level testing in 6.3. Again, the purpose of this was to confirm that actuaries were free to set different thresholds in-house – and that the model would perform well for all classes / types of business.

We note some consistent differences in model performance between short, medium and long tailed classes across all reserve deterioration thresholds considered. The cause of these differences was not clear from our analysis. We speculate that the long-tailed AUC may be lower because reserve deteriorations may emerge slower than our 3-year time horizon.

**Figure 18 – analysis of model performance - showing out of sample (OOS) AUC by reserve deterioration threshold and class type**



## 6.5. How do minor changes to the alignment score formula influence model performance?

We evaluated the performance of the model following a range of changes to the alignment score formula set out in section 4.3. The purpose of this testing was to understand the sensitivity of the alignment score feature to the various elements of the calculation.

### Table 12: impact of changing alignment metric formula on model performance

| Review area | Alignment score variation | Test AUC |
|---|---|---|
| Baseline model | | 0.711 |
| Alternative weightings between early cohorts (with potentially larger reserves) and later cohorts with fewer reserves. | Reduce weighting toward earlier cohorts by using $\sqrt{n_i}$ as the weighting instead of $n_i$ | 0.707 |
| | Use chain ladder percentage developed as the cohort weighting | 0.696 |
| Alternative approaches to capping / collaring cohorts that are significant outliers | Remove the minmax formula term – meaning that extreme values are unbounded | 0.691 |
| | Score outliers using a bucketing approach (eg which quintile they sit in) rather than capping | 0.713 |

Our testing highlights that our proposed alignment score metrics remain predictive with minor changes to the formulation of the metric. These were small relative to some of the AUC movements in section 6.2 (table 11) where the model was trained and tested on a range of data, so of limited materiality overall.

Three points to note:

- Some form of outlier capping within the formula is recommended. AUC fell when capping was removed. Intuitively this makes sense as lack of capping / collaring allows single outlier data points in a diagnostic to drive the overall score – something which a reviewing actuary would know to ignore or treat with reduced credibility if other data in the graph was behaving reasonably. We note that the effect of individual outliers is a limitation of reserving-class level AvsE analysis.
- The model exhibits some sensitivity to the choice of weighting between early (less developed) and later (more developed) cohorts. Our original hypothesis was that weight should decrease as development approaches 100%. However, we found a more statistical metric based on the number of contributing data points to be more predictive and also easier to calculate. Use of percentage development has complications where (for example) development patterns exceed 100% as well as the need for a sensible pattern to be derived.
- Using a bucketing approach to assessing deviation of data from the mean resulted in a marginally higher AUC than the base model when evaluated on the test dataset. However this was not consistently observed when evaluating other datasets, suggesting that overall there is little to choose between the approaches. Our approach of calculating (and capping) standard deviations from the mean was slightly simpler to implement and explain, so preferred from a practical perspective.

# 7. Interpretability and the human in the loop

## 7.1. Overview

The output of our multi-parameter model is a percentage score for risk of deterioration. Scores close to 0% represent little / no risk and those close to 100% represent a very high risk of future deterioration. As discussed in section 5.4, the model applies an operating threshold to determine the cut-off score that represents a risk of reserve deterioration. We set the operating threshold at 14% to maximise $F2$ and reduce the number of false negative results.

The score and classification result alone may help actuaries narrow down a large number of reserving classes to a subset that is at risk of deterioration. Within that subset, the scores themselves will help establish a priority order for further review, with higher scoring classes being reviewed first, and more marginal ones later.

However, the scoring itself will not give immediate insight into why any particular class is at risk of reserve deterioration. This insight is traditionally something that an actuary would obtain as part of the graphical review process, at the same time as identifying the risk.

We use SHAP [9] values to decompose predicted outcomes for a particular reserving class, into the underlying contribution from each of the model features, which quickly allows an actuary to:

- Identify the key features/diagnostics contributing to the modelled outcome for each entity
- Review the high scoring diagnostics to confirm materiality
- Act appropriately to reduce the deterioration risk score, or document why the results are reasonable despite the score.

This ensures that the model remains explainable and provides the reviewing actuary with similar insight to a review of the graphical diagnostics from which any model parameters are derived.

## 7.2. Calculating SHAP values

SHAP is defined as follows:

Let

$$\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$$

denote a sample of $m$ training entities – referred to as the reference set. We define a baseline prediction from the reference set as

$$\phi_0 = \frac{1}{m}\sum_{b=1}^{m} \hat{p}\left(\mathbf{x}^{(b)}\right).$$

For entity $i$, the SHAP values then sum together to satisfy the equation below

$$\hat{p}_i = \phi_0 + \sum_{j=1}^{k} \phi_{ij},$$

where $\hat{p}_i$is the modelled probability for entity $i$, $k$ is the number of features in the model, and $\phi_{ij}$ is the contribution allocated to feature $j$ for entity $i$. Where $\phi_{ij}$ is defined as:

$$\phi_{ij} = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\ (k - |S| - 1)!}{k!} [v_i(S \cup \{j\}) - v_i(S)],$$

Here:

- $N = \{1, \dots, \mathrm{k}\}$,
- $S$ is a subset of features
- $v_i(S)$ is the model prediction obtained when entity $i$'s values are supplied for the features in $S$ and the remaining features take values from the average over the reference set $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$ for those features.

## 7.3. Using SHAP to create a review dashboard based on our deterioration risk formula

Figure 19 (overleaf) provides a worked example of this for a single business class we analysed within the NAIC data. By design this could be laid out similar to a capital validation test schedule, with the proposed framework producing schedules containing the key at-risk metrics for users to analyse and comment on.

This approach could save significant time in two ways:

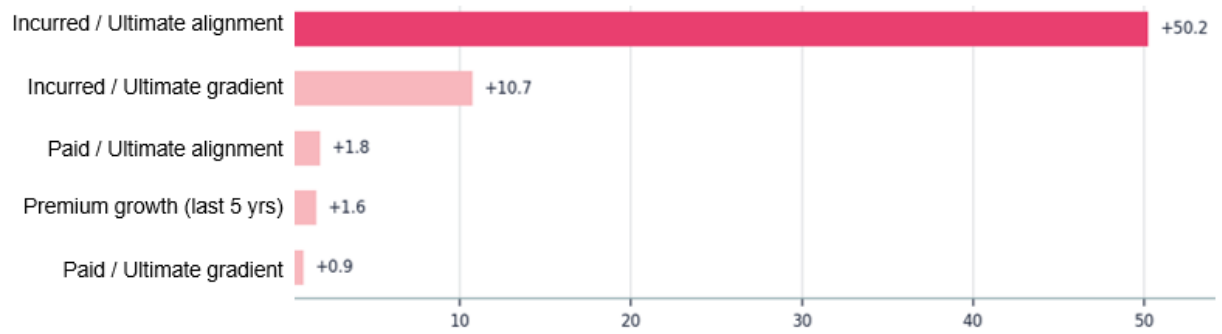
- Review could focus on classes with a reserve deterioration score beyond a certain threshold. This significantly reduces the workload for actuaries on roll-forward without relying on a single diagnostic (eg AvsE) as a guide or having to review several diagnostics for every class.
- For an at-risk class, only the diagnostics that highlight the deterioration risk need to be reviewed initially, with a clear prioritisation provided by the SHAP analysis.

**Figure 19: illustrative review dashboard**

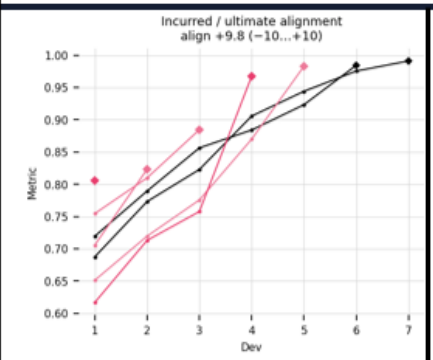
## Commercial Multiple Peril (10374)

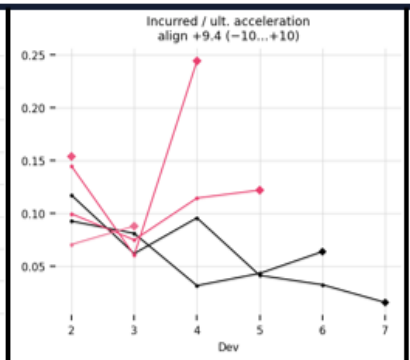
| Deterioration risk score | 88% | Flagged for review |
|---|---|---|

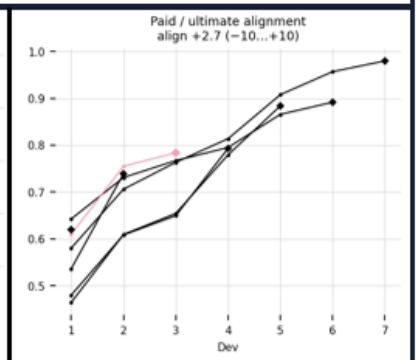
**Score drivers – Shap analysis**




**Top 3 diagnostics**




**Diagnostic commentary**

| | | |
|---|---|---|
| Clear evidence of incurred to ultimate ratio higher than historical data. No currently known company / industry reason to reserve inconsistently with past data. | Evidence that incurred to ultimate misalignment is getting worse rather than self correcting. | Misalignment behaviour not observed as strongly on paid data. Could suggest overall reserves are reasonable if paid data more reliable than incurred. |

**Discussion**

Incurred to ultimate misalignment likely requires a chain ladder model change to correct. However, it could also be driven by changes in claims process, and there is less evidence on the paid side of the data

**Next steps**

Before making a change to incurred model:

- Confirm with claims team whether there are any changes to case estimation philosophy
- Review whether projection should be based on incurred or paid based methods given inconsistency of incurred and paid diagnostics.

If continuing with incurred method, adjust projection basis to reduce incurred misalignment, and then recalculate deterioration risk score.

## 8. A two-way model that captures reserve improvements and deteriorations

During model development, we also considered how best to extend the model from predicting reserve deteriorations to also predicting future reserve improvements. This is highly desirable from an actuarial perspective, as simply flagging the risk of deteriorations, without corresponding risk of prudence, could lead to a prudent bias in reserves when aggregated across a large number of classes.

A key challenge is that our choice of a binary classification model means business groups are classified as either "expected to deteriorate" or "not expected to deteriorate".

Although the underlying problem is symmetric (ie a reserve improvement is just a negative deterioration), the model is not symmetric. "Not expected to deteriorate" is different to "expected to improve". For example, a business group or head of damage could be classified as "not expected to deteriorate" because all metrics were indicative of a best estimate reserve. In this case, the business group also wouldn't be expected to have a reserve improvement.

To extend our model to capture reserve improvements, we considered the following options:

- A single multinomial model classifying each entity as improving, stable or deteriorating.
- A separate model-development exercise for detecting improvements - resulting in a different model and feature set.
- Reuse of the selected binary modelling framework with the direction of the outcome reversed.

We consider the third approach to be a practical extension of the analysis in this paper. A reserve improvement parameter in the modelling pipeline changes the positive outcome from a material deterioration to a material movement in the opposite direction. The same feature engineering, logistic specification, chronological training design and validation controls are then applied.

Although the framework is the same, the improvement model is fitted separately and therefore has its own coefficients, probabilities and operating threshold. As discussed above, improvement risk probability is not assumed to be one minus deterioration risk probability, because an entity can fall within the stable middle range and trigger neither outcome. This approach retains a consistent and interpretable methodology without forcing improvement and deterioration to have identical relationships with the model inputs.

When the two scores are subsequently combined, each is assessed relative to its own threshold: neither threshold being exceeded produces a stable recommendation, whilst we suggest the direction furthest above its threshold is selected if both models trigger. Alternatively such cases could be flagged as uncertain or volatile.

## 9. Key conclusions and implications for reserving practitioners

Our research demonstrates that it is possible to capture the key features that actuaries look for in graphical diagnostics (such as incurred and paid to ultimate charts) using simple numerical analysis. Our incurred and paid alignment scores are shown to be predictive of reserve deteriorations and reduce the reliance on reviewing a large number of charts.

These can be calculated in a straightforward manner from the data using the following formula.

$$A(M) = \max\left[-10, \min\left(10, \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}} n_i \max\left(-5, \min\left(5, \frac{M_{i,j_i^*} - \mu_i}{\sigma_i}\right)\right)\right)\right]$$

Extending this, we have developed a logistic regression that allows us to include multiple actuarial metrics to predict reserve deteriorations. The final formula is

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -1.50 + 0.85X_{1,i} + 0.27X_{2,i} + 0.17X_{3,i} + 0.11X_{4,i} + 0.20X_{5,i}.$$

$$\begin{aligned}
\boldsymbol{\hat{p}_i} &= \text{Reserve deterioration risk score,}\\
\boldsymbol{X_{1,i}} &= \text{Incurred / ultimate alignment,}\\
\boldsymbol{X_{2,i}} &= \text{Incurred / ultimate gradient alignment,}\\
\boldsymbol{X_{3,i}} &= \text{Paid / ultimate alignment,}\\
\boldsymbol{X_{4,i}} &= \text{Paid / ultimate gradient alignment,}\\
\boldsymbol{X_{5,i}} &= \text{Premium growth (5y).}
\end{aligned}$$

The output from the formula is a percentage score that allows for ranking of classes by reserve deterioration risk.

By flagging classes with a score above 14% (the operating threshold - set based on $F2$ maximisation) we were able to:

- Flag 82% of actual future reserve deteriorations $\geq$ 5% correctly, and concurrently.
- Reduce the number of classes requiring review by 37% (assuming only the classes flagged by the model are reviewed).

This analysis suggests that our model has potential as a screening tool that would allow actuaries to quickly flag a significant proportion of a portfolio (ie 37%) as being at low risk of reserve deterioration and then prioritise the remaining classes based on the magnitude of the reserve deterioration risk score.

Using SHAP analysis, we were able to construct efficient tailored review dashboards and test schedules for problem classes that summarise the metrics that flagged a reserve deterioration risk.

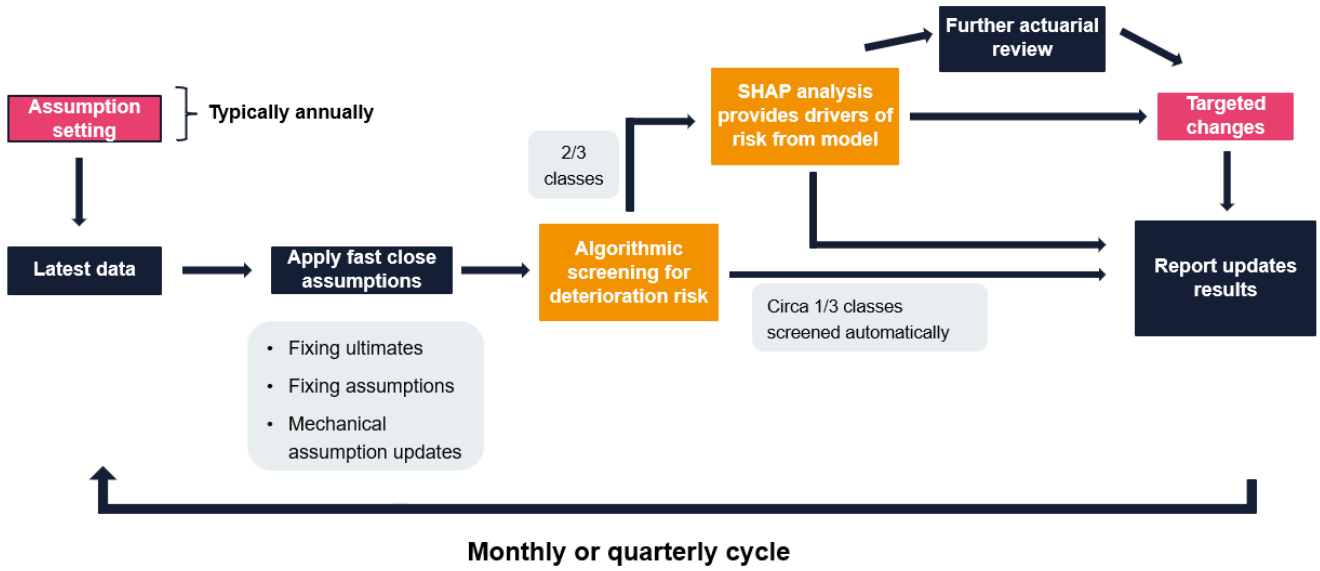
The implications for reserving actuaries are:

- The ability to use a structured screening tool to prioritise classes for review, based on reserve deterioration risk, opens the door to partial automation of reasonableness review that actuaries typically perform: after setting assumptions, on roll-forward, or when allocating reserves to a granular level for reporting.
- Being able to construct tailored dashboards highlighting just the key “at risk” diagnostics adds a further time saving to the review process by making it smarter and immediately focussing actuarial reviewers on the diagnostics that highlight the risk. This is an efficiency beyond simply flagging a reserving class because of its overall risk of deterioration.
- Incurred (and paid) alignment scoring provides a plausible supplement (or alternative) to actual vs expected analysis, being simple to calculate and avoiding many of the limitations of AvsE.

The logistic regression model is easily extendible, and the results of the evaluation and validation process in sections 5 and 6 provide a clear benchmark for iterative improvements to the model in the future. We recommend focussing on those that can flag a similarly high proportion of classes at risk (equivalent model recall), whilst further reducing the number of false positive classes requiring a more detailed actuarial review (improved precision).

At the start of this paper, in figure 1, we outlined the fast close process as typically seen within the market. Figure 20, below, summarises how this process can be adapted to be faster and smarter using the framework set out in this paper. The orange boxes highlight how the tools and techniques we propose can be easily integrated into the standard workflow.

**Figure 20: Faster close framework, incorporating approaches argued for in this paper in orange**




*An adapted fast close framework, with orange components indicating where our review framework can be deployed to improve efficiency.*

## 10. Limitations

Whilst we have confidence in the conclusions of our work, it is important to highlight the relevant limitations of our analysis:

- Our research is based on a single aggregated industry dataset, albeit over a long time period and for a wide range of companies and triangles. The annual cohort, annual development structure of the triangles differs from the more granular cohort and development intervals commonly used by UK reserving practitioners, which often have more granular cohorts and development periods than annual.
- There is a risk that our model might infer class level behaviour from the training data that it then projects into the future, but which isn't evident in other datasets. Although we have taken steps to avoid this by creating entity-disjoint testing data, some class-level behaviour can still be inferred if (for example) many or all companies for a particular reserving class behave the same way, or if classes experience similar trends across multiple eras.
- We only want to predict deteriorations due to under-reserving. The use of industry data for model training means that we implicitly capture deteriorations that are caused by other factors, for example firms increasing management margins after a good year. We attempted to minimise the effect of this by removing classes that suggested evidence of data restatement and also by validating that the mean reserve movement in the datasets was close to 0 – suggesting best estimate reserving. Nonetheless it is impossible to fully adjust for any such effects or to quantify their effect on the model.
- We used a 5% threshold for predicting reserve deteriorations. The model only predicts the risk of a deterioration > 5%. It does not attempt to predict the severity (eg a 5.1% vs 50% deterioration). There is also value in highlighting how material the deteriorations are expected to be, but this would require a more complex model formulation.
- We considered a three-year time horizon for reserve deteriorations. In practice some industry events take longer to result in material reserve deteriorations. Our analysis will not necessarily capture early indicators of reserve deteriorations that take longer than three years to manifest.

- Incurred metrics were consistently shown to be more predictive than paid metrics. We suspect this is due to wider use of incurred metrics in setting reserving assumptions. It would be useful to perform this analysis on a long-term dataset where the primary reserving methodology was based on paid claims projections. This would allow for an evaluation of whether incurred diagnostics continue to outperform paid ones, or whether there is (as we suspect) a connection between the reserving basis and the diagnostic choice.
- Our model was limited to a range of metrics compatible with aggregate market data. Other more detailed metrics available at the reserving assumption setting granularity may also prove predictive.
- There is a risk of circularity in our analysis. The metrics that we used are based on diagnostics commonly used by actuaries in graphical form to predict reserve deteriorations. If actuaries typically use incurred / ultimate ratio as a reasonableness check and respond to it by adjusting their projections, this behaviour will be reinforced in our model training and testing. Basing our training and testing on aggregate market data provides some protection from this bias because the data is transformed and aggregated before use. The three-year time horizon also limits the extent to which our model learns to predict short term volatility in ultimate claims rather than long term deteriorations that are "baked in". Nonetheless, it is possible that a contributor to the predictive power of the incurred / ultimate alignment metric is this risk of circular logic.

## 11. Recommendations for further work

We believe our work lays the foundation for the prediction of future reserve deteriorations using numerical techniques. Our analysis has identified a number of areas it would be desirable to investigate further:

- **Expanding our model to include reserve improvements:** Our initial focus was on predicting reserve deteriorations. However, it is equally important for actuaries to highlight reserving classes at risk of being over-prudent. Focussing exclusively on reserving classes at risk of deterioration could lead to overall prudence. Whilst we have set out the principles of a two-way prediction model in section 8, further work is required to evaluate the most effective model structure and parameterisation.
- **Comparison of our model performance to AvsE performance:** AvsE is the established numerical tool for evaluating which classes should be reviewed. However, differences between actual vs expected performance over a short time period are not necessarily predictive of long-term deteriorations. It would be interesting to evaluate the performance of our approach against the established industry practice. We were not able to do this (or incorporate AvsE into our model) as it is not a KPI within the Schedule P data we were using.
- **Improving the performance of the existing model:** The existing model performs well, but there is likely scope to retain the same 82% recall, whilst reducing the number of classes that the model flags to achieve it. Some suggested next steps include:
    - Expansion of the training data to a wider pool than the Schedule P returns, with a particular emphasis on including larger and more granular triangles aligned to the typical granularity at which actuaries reserve.
    - Consideration of other KPIs that may boost predictive power, eg analysis of claim frequencies and severities, development patterns and inflationary trends.
- **Testing on broader datasets:** Our training and testing data were limited to the Schedule P returns. We recommend evaluating the performance of the model on a range of company data and other industry data to investigate how broadly applicable the model is.

# 12. References


[1] ROC Working Party, *Towards the Optimal Reserving Process: The Fast Close Process*, March 2015, British Actuarial Journal 20(3):491-511 DOI:10.1017/S1357321715000021

[2] Balester J C, Kirschner G S, *Structured Tools to Help Organize One's Thinking When Performing or Reviewing a Reserve Analysis*, November 2013, Casualty Actuarial Society website, Available at https://forum.casact.org/article/166179-structured-tools-to-help-organize-one-s-thinking-when-performing-or-reviewing-a-reserve-analysis (accessed 30 August 2026)

[3] Diffey W, Hobern L, Lauder A, et al. *Your reserves may be best estimate, but are they valid?,* British Actuarial Journal. 2022;27:e2. doi:10.1017/S1357321721000179

[4] LCP, *Global reserving and transformation review*, July 2024, LCP website, Available at: https://insights.lcp.com/rs/032-PAO-331/images/LCP-Global-Reserving-and-Transformation-Review-07-2024.pdf (accessed 30 August 2026)

[5] TORP Working Party, *Towards the Optimal Reserving Process: Actual versus expected techniques*. IFoA website, September 2013. Available at: https://www.actuaries.org.uk/system/files/documents/pdf/e8-torp-giro-2013.pdf (accessed 30 August 2026).

[6] Hanley JA, McNeil BJ. *The meaning and use of the area under a receiver operating characteristic (ROC) curve*. Radiology. 1982;143(1):29–36. doi:10.1148/radiology.143.1.7063747

[7] Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 1950;78(1):1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[8] van Rijsbergen CJ. *Information Retrieval*. 2nd ed. Butterworth-Heinemann; 1979.

[9] Lundberg SM, Lee SI. *A unified approach to interpreting model predictions*, Advances in Neural Information Processing Systems. Vol 30. Curran Associates, Inc.; 2017:4765–4774. doi:10.5555/3295222.3295230