Harmonising data: The art of validation and management

Eoin King
Arushi Mittal

Introduction

Data is of significant importance for insurance and reinsurance companies, given it is used for various purposes such as regulatory and financial reporting, as a basis for public disclosures and as an input into strategic decision making (including portfolio acquisitions, launch of new channels, pricing new products etc.), amongst others.

Hence, it is crucial for companies to ensure the accuracy, consistency and reliability of data before it is used. This is also important to reduce the risk exposures for insurers, especially operational risk which may stem from a lack of data quality or inadequate data validation and management systems.

With the recent implementation of International Financial Reporting Standard (IFRS) 17, companies need a more detailed level of granularity of data, which necessitates the importance of having strong data controls to ensure high-quality results. The need for data validation is further augmented as the companies will aim to use the same data across multiple reporting bases and will need to ensure consistency across the same. Any inconsistencies may stem from the data being used at different levels of granularity across different teams and systems or differences in the date when the data extract was taken etc. Further, with the rise in use of artificial intelligence (AI) and machine learning as well as new tools and languages, such as R/Python, Power Query etc., across the insurance sector, firms may now look to use these technologies to enhance data validation checks and quality controls.

A recent newsletter[1], issued by the Central Bank of Ireland (CBI) in September 2023, highlighted that a number of insurance firms have identified operational errors within reserving which were driven by weak controls in manual processes (which may be required due to data issues). The CBI elaborated that high-level checks such as analysis of change were not sufficient to surface the errors, and only more in-depth validation led to identification of the issues. The CBI further stressed the importance of having effective systems, controls and documentation for the purpose of valuations of assets and liabilities. In addition, the same newsletter called for validation of external data in the context of non-life reserving.

Our earlier briefing note, "Structuring an Effective Model Risk Management and Validation Framework,"[2] provides a comprehensive overview of model validation framework, including inputs, calculations and results output validation.

This briefing note delves into the importance of data validation and the current challenges faced within insurance companies. It further covers how companies can tackle data validation and implement data validation frameworks to address concerns regarding data quality and its governance.

WHY DO WE NEED IT?

Data validation can have wide-ranging benefits to firms, a few of which we discuss below.

End uses of data

Implementation of data validation mechanisms and governance of key input data aids insurers to meet requirements relevant to various end use cases, for example regulatory and financial reporting under various regimes such as Solvency II and IFRS 17, where data is used for experience studies and assumption setting, cash flow modelling, asset liability management and creating model points for pricing new products etc.

---

1 Gibney, C. & Bradley, M. (September 2023). Reflections on Reserving. CBI Insurance Newsletter. Retrieved 13 June 2024 from https://www.centralbank.ie/docs/default-source/regulation/industry-market-sectors/insurance-reinsurance/solvency-ii/communications/insurance-quarterly-news/the-insurance-quarterly-september-2023.pdf?sfvrsn=45e49c1d_5.

2 McGinley, D., Gleeson, C., McIlvanna, M., & Stack, E. Structuring an Effective Model Risk Management and Validation Framework. Milliman Research Report. Retrieved 13 June 2024 from https://www.milliman.com/-/media/milliman/pdfs/2021-articles/4-6-21-model-validation-considerations.ashx.

Strategic need

Working with accurate, reliable and consistent data, along with a good data governance framework, allows insurers to gain more confidence in their results and use this data for various strategic needs, such as for launching new products, producing management information, deciding on reinsurance arrangements, analysing mergers and acquisitions, making portfolio transfers etc., where reliable data is crucial to correctly price the products or deals to gain competitive advantage and avoid any subsequent legal suits.

Regulatory and audit requirements

The European Insurance and Occupational Pensions Authority (EIOPA) addresses data validation as part of the Solvency II Delegated Acts in Articles 19-21.[3] It requires companies to have internal processes and procedures to ensure data is appropriate, complete and accurately used for the calculation of technical provisions. It further requires companies to have internal control processes for independent review and validation of data on a regular basis.

Auditors are also increasingly focussing on quality of data, stressing the need for adequate and reliable data controls to ensure accuracy. Having clean and transparent data validation checks and controls along with appropriate documentation will help to meet the audit requirements and avoid any material observations by auditors.

In the following sections, we discuss the key challenges regarding data that insurers may face and how they can be addressed.

ISSUES CONCERNING DATA MANAGEMENT

Here are some key issues that insurance firms may face in respect of managing and validating data.

Incomplete or missing information

Insurers may receive incomplete or missing data from internal teams or from external third-party providers. This may stem from systemic issues in the data collection processes and systems, loss of files, data entry errors or omissions or incomplete data from external sources.

Working with missing information may also lead to model runs falling over if any expected fields are not populated, leading to increased run times and operational inefficiencies. Incomplete data or omissions in data, if not identified, may lead to an overstatement or understatement of results by firms, which may reduce the desired level of confidence the firm wants over the key results and lead to inaccurate reporting. The impact may be increased if the results from errors in data are rolled forward to subsequent periods.

Multiple sources of data

Actuaries often work with data from multiple sources, both internal and external, e.g., for data from a policy administration system, claims database, finance/investment systems, reinsurers, asset managers and product literature. These data sources can be tricky to navigate if inconsistencies exist across them. The inconsistencies may arise due to differences in data structures, formats or conventions being used, human errors in data entry, data migration errors or differences in the timing of data across the multiple sources.

These issues become even more difficult to identify when different sources of data are used for different reporting bases. Firms may incur substantial time and resources to reconcile data from multiple sources before it can be used for actuarial model runs and to reconcile the results produced on various reporting bases. These issues can potentially further increase if the number of sources or the volume of data increases over time.

Erroneous data

Errors in data may have occurred due to data entry errors, e.g., due to typos, incorrect data input or misunderstanding of data requirements. Errors may also arise from system glitches or malfunctions that may cause data to be processed or recorded incorrectly.

Unidentified errors in data can generally lead to actuaries using a poor quality data as part of actuarial processes. Substantial amounts of time and resources may have to be dedicated towards manually correcting results as well as rerunning processes if the data errors are spotted at later stages.

Lack of quality data may lead to the insurer needing to hold more capital to address the data quality issues, especially if the regulator has expressed concern in this regard.

Challenges with data structures

Inefficiencies in how data is organised and stored can create complexities in managing the data and higher turnaround times. A lack of understanding of the data, inconsistent formats, naming conventions or folder structures or inefficient data hierarchies can impede accurate modelling. Lack of validation controls in respect of data structures will further increase the exposure to operational risks across various areas and decrease traceability and efficiencies.

ADDRESSING DATA ISSUES FACED BY INSURERS

Data validation framework

Insurers can address various issues or challenges related to input data by following good governance practices. This entails having or setting up an appropriate data validation framework. This will enable a clear and comprehensive definition of steps to be followed along with good practices and assignment of roles and responsibilities.

One of the potential examples of a data validation framework is shown in Figure 1 and discussed below.

FIGURE 1: DATA VALIDATION FRAMEWORK EXAMPLE
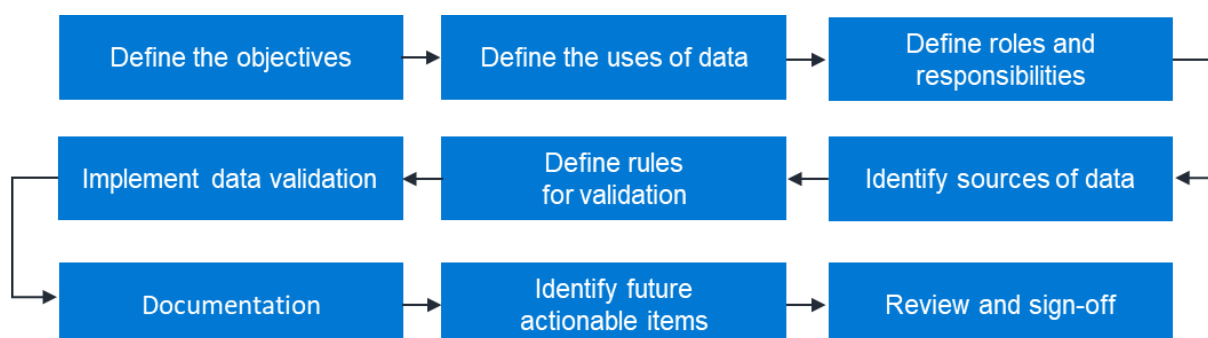
1. Define the objectives

    The key objectives of the framework could be to build, maintain and monitor robust data management processes to ensure that data is accurate, reliable and complete to be used for various end uses as defined in the next step.

    Insurers may add other objectives to the above which may be specific to their company requirements and previously identified issues regarding data.

2. Define the uses of data

    Companies should clearly define how they intend to use the data. This could include financial reporting and analyses of data for experience studies, assumption setting, risk assessment and pricing, among other uses.

3. Define roles and responsibilities

    Companies should define a clear assignment of roles and responsibilities for individuals. This will help in setting clear expectations for team members, minimise conflict and aid in improving operational efficiency. It will also reduce any undesired duplication of work. Assigning responsibilities may also boost the sense of ownership over a specific task and may help improve productivity through structured planning.

4. Identify sources of data

    It is important to identify and list all sources of data used for any project that a company may undertake. This can help companies carry out reliability assessments for each source of data, identify which sources of data to use and also to identify where the company may be exposed to the risk of relatively poor-quality data, which may require a higher number of checks and controls to be performed.

5. Define rules for validation

    The next step of the framework involves laying down the rules for data validation. It can entail the following steps:

    - Define the types of checks and validation rules
    - Define internal materiality/tolerance limits
    - Describe approach to correcting data and populating any missing information
    - Define when and how to apply an expert judgement

6. Implement data validation

    This can entail the following steps.

    - Carry out checks and controls, which are further discussed in the next section below
    - Error resolution
        - Apply materiality limits
        - Apply rules for correcting data and populating any missing information
        - Apply expert judgements where needed
    - Identify action items for future

7. Documentation

    A data validation report, clear and comprehensive documentation which clearly lists checks and controls as well as errors in data and expert judgements applied, is essential. It enhances the reliability of data used in actuarial models and also allows stakeholders to have confidence in the results produced from the data. Data validation reports may also help in meeting regulatory and audit requirements, providing evidence of data controls.

    This also helps reduce any key person dependencies by enabling reproducibility and can also help entities identify areas where additional validation might be needed in the future.

8. Review and sign-off

    Review and sign-off ensures there is appropriate governance and that work performed is of a high quality and free from errors. It is also critical for risk management and accountability.

9. Identify future actionable items

    It is important to note any topics or issues that can be worked upon or improved in the future. This allows insurers to prioritise their tasks and address any actionable points as necessary.

DATA VALIDATION CHECKS

Depending on the nature of data and the use cases, different types of checks and controls can be employed to ensure a suitable validation process is used.

We have developed a comprehensive in-house checklist of various data validation checks and controls that can be used by clients as part of their own data validation processes to ensure reliability and accuracy of data or for independent reviews on data validation and controls.

Broadly, the checks to address data issues can be split into the following categories:

Preprocessing checks

These checks are performed on the raw data extract before model point files are generated. These checks allow for the identification and rectification of potential inconsistencies and errors with the input data.

Examples of preprocessing checks that can be performed by firms include checks on:

- The number of files expected to be used
- Any manual data processes or adjustments made
- File formats (e.g., Excel, CSV or other text files) and that the formats of fields used within the files are as expected
- Identifying any loss of data, for example checks on key total values like premium and benefit values (sum assured/annuity amount/unit value), after any transformation or grouping process, checking them against totals provided from policyholder administration teams etc.

Periodic consistency checks

Periodic checks involve comparing data from the current period against the data used at the previous valuation period. These checks can be implemented for input data for both policyholder data and asset data files.

Examples of the checks included in this category are:

- Movements in the number of policies since the previous period (allowing for new business and exits over the period).
- Changes in policy details over the said period. This can include checks around static fields (which are not expected to change from one period to the next) and dynamic fields (which might change more frequently). Some examples of static fields include the start date of the policy, retirement age, commission rate, sales channel etc. Some examples of dynamic fields include duration in force, unit price etc.

Generic checks

These checks involve general checks on logic and checks on consistency for various input data. They can include checks to determine the validity of the values in certain fields as well as checks on the records which fail validation rules but are within agreed internal tolerance limits and align with the entity's risk framework.

Examples of these checks can include:

- Maximum start date is less than or equal to valuation date
- Dates for paid and incurred claims should not be in the future
- Fields such as sum assured, annuity amount and fund value are greater than zero

External consistency checks

This category of checks entails checks to ensure that data is consistent across various data sources used by a firm.

These checks may include:

- Checks on floors and ceilings of fields to validate that the fields lie within the ranges specified in the product literature. For example, floors and ceilings on fields such as premiums, management charge percentages, sum assured amounts etc.
- Checks on new business model points against the company's management information reports. For example, new business data such as number of new business policies, premium amounts etc.

TOOLS AND TECHNOLOGIES TO CONSIDER FOR DATA VALIDATION

Centralised data warehouses

Firms can maintain data lakes, centralised databases or warehouses for all their data. Data lakes allow the storage of raw data in its native format, and offer flexibility, scalability and the ability to handle a vast diversity of data types, making data lakes an essential component of big data and analytics strategies. The availability of diverse datasets in one place fosters innovation and deeper insights through advanced analytics and machine learning. However, as data lakes work with raw data, further validation may be needed to provide cleansed, refined and aggregated data. On the other hand, centralised databases provide a single source of the truth of data, which reduces the chances of discrepancies and inconsistencies in the data. It also enforces standardisation of data formats, making it easier to validate data. In a centralised data warehouse, errors can be detected and corrected in a single location, rather than at multiple places. It may also have built-in data quality checks, ensuring that data is validated upon entry into the system.

Structured Query Language (SQL)

While SQL is commonly associated with querying and manipulating data, it also offers powerful features and techniques to validate data, making it a valuable tool for maintaining data integrity and accuracy. SQL can handle large datasets in comparison to tools like Excel. In Excel it is relatively simple to alter data, but SQL, as a database server, reduces the chances of making any accidental changes to the data. Changes need queries to be executed to come through to the database. SQL also offers tools and techniques that streamline the validation process by improving query performance. Efficient queries can handle larger volumes of data and make the data validation process more scalable while also making the process faster and more efficient. SQL further allows for easy replication of checks and controls on new sets of data, improving the turnaround times for processes.

R/Python

R and Python are open source programmes which can be very cost-effective and can be used to validate large volumes of data. These programmes provide the functionality to implement a robust data validation framework as well as offering tools for data visualisation.

These programmes can be used to create functions or routines that assess whether input data meets specific criteria, as well as to provide functionality to integrate complex data structures with existing codebases. Python libraries support various data formats and sources and provide an easy way to validate data against expectations and alert the user when expectations are breached.

Power Query

Power Query is a robust tool which enables users to perform data validation and checks by applying rules or conditions to the data, and aids in identifying and correcting data quality issues. It supports a broad array of data sources, including files and databases, making it easy to analyse data across various sources. It has an extensive library of built-in data transformation functions, which allow it to perform complex data manipulation tasks.

Artificial intelligence (AI) and machine learning (ML)

AI and ML techniques can have potential future uses by allowing users to automate their data validation processes. They can also allow for algorithms to learn from data and improve over time, making the validation more accurate to the data concerned. They can predict future trends based on historical data, which can be useful for data validation processes. AI and ML can also be used to detect anomalies and outliers in data, which can be used for detecting errors and inconsistencies in data. For example, generative AI can be used to simulate corner cases that might not be present in the available data but could occur in real-world scenarios, which can help generate validation rules that can handle these cases. It can also be used to simulate different types of errors or corruptions that might occur in data, which can be used by firms to validate that their existing checks and controls address this issue.

Data visualisation tools

Data visualisation tools enhance the ways in which users interpret and analyse data and can prove to be instrumental in optimising and streamlining data validation processes. They can enable users to identify and understand trends, data patterns and correlations, comparing datasets and identifying outliers and anomalies as well as comprehensively communicating validation results. Data validation may become more efficient and intuitive by leveraging data visualisation techniques and tools such as Tableau, Power BI etc.

Milliman Mind

The web-based Milliman Mind tool can convert Excel spreadsheets into powerful models with all the features and auditability of more expensive "black box" systems. Milliman Mind can be employed to establish robust data validation processes by firms. It allows firms to streamline and simplify existing processes in Excel with enhanced security, auditability and performance. It can transform an Excel-based process into a cloud-based all-in-one solution. which reduces the amount of manual intervention needed, reducing the risk of errors and broken links, slow runtimes and lack of scalability. It further allows for an automated production of an audit trail, a workflow and dashboards.

CONCLUSION

The significance of data validation in the insurance industry cannot be overstated, especially in light of the increasing regulatory requirements and the growing complexity of data sources and structures. Firms with legacy data validation processes may face operational inefficiencies if they have not taken the time to automate, improve or continuously monitor their data processes.

In this briefing note, we have highlighted the importance of data validation and further discussed challenges faced by insurers in data management. To address these challenges, we have included a specimen of a data validation framework, including checks and techniques that firms can employ for a more robust and comprehensive data validation process.

Milliman

Milliman is among the world's largest providers of actuarial, risk management, and technology solutions. Our consulting and advanced analytics capabilities encompass healthcare, property & casualty insurance, life insurance and financial services, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT
Eoin King
eoin.king@milliman.com

Arushi Mittal
arushi.mittal@milliman.com