

# Modelling a medical expenses portfolio

Supplementing traditional insurance approaches  
by using clinical data algorithms

Joanne Buckle, FIA  
Nicola Biscaglia, ISOA  
Jose Silveiro, IA



## Use of more sophisticated techniques can transform the way insurers predict and manage claims and keep premiums affordable.

Medical inflation is a significant challenge when managing health insurance portfolios. While there are different ways to think about the components of medical inflation, fundamental capabilities for managing medical expense books include front-end medical underwriting to manage the health costs of the covered population and deter anti-selection, supply chain management and member engagement programmes to keep the covered population as healthy as possible and manage selective lapsing. Unlike other insurance products, such as motor and household, there tends to be less discussion among actuaries in health insurance about advanced claims prediction and pricing techniques. This is partly because in some markets pricing for health insurance is heavily regulated. However, in many European markets, pricing for medical expenses follows regulation similar to any other insurance line and statistical techniques such as generalised linear models (GLMs) may be used to determine frequency and severity models. However, there is limited understanding of the provider data flows that distort the ability to define a “claim” consistently over time and limited use of clinical data to enable the more advanced insights into future claims. In this paper we discuss why and how insurers can apply different methods for predicting claims and lapses in health insurance and how to overcome some of the barriers.

## Medical insurance data

### MAKING SENSE OF MESSY AND COMPLEX DATA

When compared with other insurance lines, medical expense claims data appears initially overwhelming. Depending on the benefit package, the majority of covered lives may have multiple claims a year and may have many hundreds of invoices relating to different services from a plethora of diverse providers, doctors, pharmacists, hospitals, clinics, primary care facilities etc. The concept of a “claim” is not clear-cut—is it an invoice? An “episode of care”? All services that happen within a specified time period?

Other challenges with health claims data include different provider contracting arrangements, which lead to different billing patterns (case rates vs. fee-for-service billing, for example). In addition there can be data quality issues around the integrity of clinical coding if it is not necessary for payment. It is extremely important to understand the provenance of the data, carry out clinical audit checks and ensure the input of medical and coding professionals.

These definitional and operational issues make traditional frequency/severity modelling of “claims” challenging. To make sense of the data, more advanced insurers use several different types of grouping algorithms, either homegrown or, increasingly, standardised commercially available algorithms. These “groupers” fall into different categories:

**Service line groupers:** Services are grouped according to type of facility or type of benefit. For example, grouping all types of outpatient scans together into one line of data, with an aggregate utilisation and average cost of services, or grouping all types of medical or surgical inpatient admission together with a utilisation rate of admissions, average length of stay and average cost per admission. Service line groupers, such as the Milliman Health Cost Guidelines ([Health Cost Guidelines Suite | Milliman | Worldwide](#)) are a necessary building block for many types of claims-related analytics.

**Clinical groupers:** These grouping algorithms encompass episode groupers, where all services are grouped into one “episode of care” related to one condition or illness, which may encompass many different elapsed time periods, from days to months. A typical example might be to group preadmission testing, an inpatient admission and most discharge follow-up, all related to one treatment, into an episode of care. Other clinical groupers include population segmentation tools, where members are “tagged” as belonging to any one of a number of chronic condition groups, and a hierarchy is imposed in the event that members have more than one qualifying condition. The purpose of the chronic disease tag is not to stigmatise patients as poor ongoing risks, but to understand which patient groups are driving claims cost trend, their specific comorbidities and costs and how their claims cost is likely to develop over time. Such population segmentation also helps when considering future product, benefit and efficient care management design. To use clinical groupers, minimum diagnosis coding such as ICD codes must be present on the claim, along with details of the services delivered.

An example of a clinical grouper is the Milliman Chronic Conditions Hierarchical Groups (CCHGs), where members are categorised into one of approximately 30 different health status groups, which typically have far more explanatory power than using age and sex, which are traditionally the most important factors used in a predictive model of claims costs. Clinical groupers can also encompass diagnostic groupers where similar treatments are categorised together for payment, as they consume similar levels of resources.

**Risk-scoring algorithms:** More sophisticated and powerful tools for the predictive modelling of claims are risk-scoring algorithms, such as the Milliman Advanced Risk Adjusters (MARA Milliman Advanced Risk Adjusters (MARA) | Milliman | US). These algorithms use historical claims data with diagnosis and treatment codes to assign risk scores to members, allowing insurers to identify the average risk of the portfolio over time. They also rank members according to metrics such as rising risk scores, which may indicate a need for clinical intervention, for example to prevent unnecessary admissions.

## Projecting claims costs and lapses

### TYPICAL VARIABLES

Use of clinical and grouping algorithms allows data to be modelled in manageable and easily explainable categories. Usually projections are carried out within broad service line categories, for example inpatient admissions, outpatient services, physician services (where they are billed separately). It is common to use “frequency and severity” modelling, but recast as “utilisation and unit cost of service” modelling, perhaps with risk scores or population health status or clinical status group as an explanatory variable. The key difference is the modelling of clinical/healthcare “services” rather than “claims.”

Other variables that tend to have strong predictive power include age and sex, geography (both for utilisation and unit cost), duration since underwriting (if medical underwriting has been applied), group size, distribution channel and hospital network or other policy/benefit types and a time variable to capture underlying medical inflation.

There are significant benefits to a multivariate approach over the univariate approach. Being able to isolate medical inflation and its impact on both the use of healthcare services and the cost is important and often small shifts in demographics and health status can have significant effects in overall claims costs and obscure true medical inflation. This makes claims forecasting difficult, especially given the disruption in claims frequency and mix due to COVID-19 from 2020 to 2023.

GLMs have been used in medical insurance for claims and lapse forecasting, as well as pricing in some countries for many years and are now being supplemented by machine learning approaches, which automate much of the (sometimes lengthy)

GLM process. We present below a case study aimed at providing Italian insurance companies with a benchmark for the claims costs of health benefits, enabling them to optimise their ratemaking. The case study is based on the Italian health market data analysed through a multivariate approach. The goal is to provide health insurance companies with a helpful benchmark to improve their healthcare risk pricing models. This helps to mitigate the lack of internal health data at a time where demand for health insurance products is growing rapidly due to the challenges faced by the National Public Health System.

Around 7 million lives in health data coming from different insurance portfolios were integrated into a single dataset using data standardisation processes. Through effective preprocessing and advanced clustering data techniques, multivariate GLMs were implemented to develop a model to assess the underwriting risks associated with both individual and group health benefits.

### BENEFITS GROUP, CLUSTERING AND MODEL IMPLEMENTATION

These benefits have been grouped into service line categories, such as “hospitalisation, surgery and day-hospital” and “major Interventions,” which are the focus of the analysis. After completing an exhaustive exploratory data analysis, the selected variables were clustered based on similar risk predictions, serving as predictors for actuarial models.

Using specific ratemaking software that employs machine learning and predictive analysis to simultaneously execute several GLMs, the utilisation rate and average cost models for the grouped benefits were obtained and used to generate index variables to estimate the claims cost for different portfolio compositions.

Unlike the univariate approach, the risk relativities coming from these predictive models reflect the risk relationships, particularly the linear risk correlations, both with the target variable and between the identified predictors (e.g., territory, gender, number of household members, insured age).

### BENEFITS, DISADVANTAGES AND RESULTS

The implementation of these models has led to the development of a benchmark quotation system to evaluate individual and portfolio risks, tailored to each selected risk profile. This tool provides useful and effective support to insurance companies for pricing and monitoring risk with its effectiveness substantially enhanced by the incorporation of variables identified as highly significant within the Italian context, such as the gender-age bivariate.

The tool's reliability and validity are supported through a multivariate approach, which provides more robust and reliable results, as well as a deeper understanding of the studied phenomenon compared to a univariate or bivariate approach.

## Next steps

Pricing maturation can materially affect the performance of medical expense portfolios, but to use sophisticated actuarial methodologies it is necessary to transform the data into manageable categories. A combination of algorithms can benefit the useability of the models and aid interpretation for pricing, claims cost prediction and portfolio projection purposes. There are significant benefits of a multivariate approach over a simple univariate approach and, in some markets, the use of more sophisticated models has led to increases in profit margins for the first movers. Even in markets where pricing is significantly constrained, or rating factors are limited, understanding risk factors that lead to claims has advantages in portfolio management. Understanding and leveraging any level of clinical information on the claims data in a timely manner is critical to gain insights into future claims, understand the drivers of medical inflation and implement pricing changes in an agile manner.

---

## Solutions for a world at risk™

Milliman leverages deep expertise, actuarial rigor, and advanced technology to develop solutions for a world at risk. We help clients in the public and private sectors navigate urgent, complex challenges—from extreme weather and market volatility to financial insecurity and rising health costs—so they can meet their business, financial, and social objectives. Our solutions encompass insurance, financial services, healthcare, life sciences, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

[milliman.com](https://milliman.com)



### CONTACT

Joanne Buckle  
[joanne.buckle@milliman.com](mailto:joanne.buckle@milliman.com)

Jose Silveiro  
[jose.silveiro@milliman.com](mailto:jose.silveiro@milliman.com)

Nicola Biscaglia  
[nicola.biscaglia@milliman.com](mailto:nicola.biscaglia@milliman.com)