Jurnal Statistika Dan Komputasi (STATKOM), Vol. 5 No. 1 (June 2026)
ISSN : 2963-038X (print), ISSN : 2963-0398
DOI : https://doi.org/10.32665/statkom.v5i1.6426

Aggregate Loss Modeling for Renewal Gross Premium Estimation in Group Health Insurance

Using Zero-Inflated Poisson-Lindley and Mixture Gamma-Rayleigh Distribution

Jessica Sie[1], Achmad Zanbar Soleh[2], Lienda Noviyanti[3]

[1,2,3]Actuarial Science Study Program, Universitas Padjadjaran
E-mail: jessicasie168@gmail.com

Submitted: April 7, 2026 ***Revised:*** June 18, 2026 ***Accepted:*** June 23, 2026

Abstract

Background: *Group health insurance policy renewals require insurers to re-evaluate premiums based on historical claim experience. Medical inflation heightens uncertainty by driving up future healthcare costs.*
Objective: *This study estimates the renewal gross premium for a group health insurance portfolio using an aggregate loss risk modeling approach.*
Methods: *Historical claim data covered 912 insured individuals, with 102 claimants and 174 claims recorded between July 2024 and June 2025, drawn from a corporate client of an Indonesian insurance company. Claim frequency was modeled using the Zero-Inflated Poisson–Lindley (ZIPL) distribution to handle excess zeros and overdispersion, while claim severity was modeled with the Mixture Gamma–Rayleigh Distribution (MGRD) to capture medical cost heterogeneity. Parameters were estimated via Maximum Likelihood Estimation, and a 95% confidence interval for expected severity was derived using Chebyshev's inequality. The model assumes a 16.2% medical inflation rate, a 6.5% annual interest rate, and an 18% loading factor.*
Results: *At a 95% confidence level, the estimated renewal gross premium ranges from IDR 2.053 billion to IDR 4.352 billion. Sensitivity analysis confirms that premiums rise with increasing medical inflation.*
Conclusion: *The interval-based approach provides a statistically grounded numerical basis for underwriting decisions and negotiations in group health insurance renewals.*

Keywords : *Group health insurance, Medical Inflation, Zero-Inflated Poisson-Lindley, Mixture Gamma-Rayleigh Distribution, Gross Premium.*

Abstrak

Latar Belakang: Pembaruan polis asuransi kesehatan kelompok mengharuskan penanggung mengevaluasi ulang premi berdasarkan pengalaman klaim historis. Inflasi medis meningkatkan ketidakpastian biaya layanan kesehatan di masa depan.
Tujuan: Penelitian ini mengestimasi premi kotor lanjutan portofolio asuransi kesehatan kelompok menggunakan pendekatan pemodelan risiko kerugian agregat.
Metode: Data klaim historis mencakup 912 tertanggung, dengan 102 pemohon klaim dan 174 klaim tercatat antara Juli 2024 hingga Juni 2025, dari klien korporat perusahaan asuransi Indonesia. Frekuensi klaim dimodelkan dengan distribusi *Zero-Inflated Poisson–Lindley* (ZIPL) untuk menangani kelebihan nol dan overdispersi, sedangkan severitas klaim dimodelkan dengan *Mixture Gamma–Rayleigh Distribution* (MGRD) untuk menangkap heterogenitas biaya medis. Parameter diestimasi menggunakan *Maximum Likelihood Estimation*, dan interval kepercayaan 95% untuk ekspektasi severitas klaim diperoleh dari pertidaksamaan Chebyshev. Model mengasumsikan inflasi medis 16,2%, suku bunga 6,5%, dan faktor *loading* 18%.
Hasil: Pada tingkat kepercayaan 95%, premi kotor lanjutan berkisar antara IDR 2,053 miliar hingga IDR 4,352 miliar. Analisis sensitivitas menunjukkan bahwa premi kotor meningkat seiring naiknya inflasi medis.
Kesimpulan: Pendekatan berbasis interval memberikan landasan numerik yang kokoh secara statistik untuk keputusan underwriting dan negosiasi lanjutan asuransi kesehatan kelompok.

Kata kunci: Asuransi Kesehatan Kumpulan, Inflasi Medis, *Zero-Inflated Poisson-Lindley, Mixture Gamma-Rayleigh Distribution,* Premi Kotor.

INTRODUCTION

Insurance is a risk-sharing tool, allowing individuals from a similar group to contribute to a shared fund to reduce potential financial losses. Generally, the insurance sector is divided into two primary categories, life insurance and general insurance. General insurance, commonly known as non-life insurance, includes all indemnity-related plans that are not categorized under life contingencies. Unlike life insurance, where claim amounts and timing can often be anticipated using life tables, general insurance is characterized by uncertainty in both claim occurrence and claim size. Under general insurance contracts, policyholders may submit multiple claims within a single coverage period, and insurance contracts are typically renewed periodically. As a result, insurers face uncertainty not only in whether claims will occur, but also in the total cost that must be paid, which is crucial for insurance premium determination (Selvakumar et al., 2023). Underestimating future losses may result in inadequate premiums and financial strain for insurers, whereas overestimating losses may reduce premium competitiveness and increase the likelihood of policyholder attrition during policy renewal.

A premium is the price an insurer sets in exchange for taking on the risk that an insured party would otherwise bear alone (Manurung et al., 2025). In practice, the financial exposure an insurer carries comes from the claims it pays out, and that exposure hinges on two things together: how many claims are filed and how costly each one is. Pinning down the distributional behavior of both claim occurrence and claim severity is therefore central to estimating expected losses and pricing premiums with any confidence (Selvakumar et al., 2023). Group health insurance is one setting where these characteristics are especially visible.

Group health insurance extends coverage to a defined group, usually a company's employees, with the employer acting as the policyholder. A typical plan bundles outpatient and inpatient care together with dental and maternity benefits. What makes pricing distinctive here is medical inflation, the steady rise in healthcare costs over time, which feeds directly into the claims an insurer must pay (Khandeka & Shaikh, 2022). Since these policies renew on a regular cycle, insurers have to revise renewal premiums to keep pace with expected growth in healthcare spending; ignoring medical inflation risks setting premiums that fall short of future claim costs. In Indonesia, the pressure is acute: medical cost inflation in employer-sponsored plans runs ahead of general inflation, pushed along by an aging population, high-cost diseases, and climbing treatment prices (Prastyo & Gani, 2025). The cost structure of the claims themselves adds another layer. A routine outpatient consultation produces a modest claim, whereas a hospitalization, surgery, or advanced procedure can be far larger, so a single portfolio effectively contains several severity subpopulations generated by quite different treatment pathways. Taken together, this wide dispersion in claim size means that determining a renewal premium calls for both an inflation adjustment and a careful model of claim severity.

A standard route to premium calculation is the aggregate loss model, where total losses emerge from the interplay of claim frequency and claim severity (Klugman et al., 2019). These two components are the backbone of loss modeling and pricing, especially in experience-based pricing frameworks (Oh et al., 2020). Kusumadewi et al. (2022) priced fisherman micro-insurance with an aggregate risk model, pairing a Poisson distribution for frequency with an Exponential distribution for severity, while Manurung et al. (2025) turned to a Negative Binomial distribution for frequency and a Log-logistic distribution for severity in pricing motor vehicle insurance. A different approach was adopted by Lestari (2022), who used a Monte Carlo simulation to predict insurance claims. Simulation, however, does not explicitly model how claim frequency and claim severity interact. This study, therefore, adopts an aggregate loss framework that captures both components jointly, evaluating several candidate distributions for each and selecting the best fit through goodness-of-fit tests and information criteria.

On the frequency side, the dataset covers every insured individual in the portfolio, claimants and non-claimants alike. Observing the full population makes excess zeros a prominent feature, one that ordinary count models handle poorly. Zero-inflated models address it by splitting the zeros into two sources: insured individuals who are structurally unable to generate a claim, and those who could claim but happened not to during the period (Qazvini, 2019). In health insurance, this split is natural, since healthier or low-utilization members simply record nothing over a given year. The Zero-Inflated Poisson (ZIP) is the usual first choice for excess zeros, but it assumes equidispersion and performs poorly when the data are overdispersed. The Zero-Inflated Negative Binomial (ZINB) addresses this with an extra dispersion parameter, though at the price of a more complex model.

This motivates the Zero-Inflated Poisson-Lindley (ZIPL), a more parsimonious alternative. Its base, the Poisson-Lindley distribution, is a Poisson mixture with a Lindley mixing distribution and is overdispersed for every parameter value, meaning $Var(Y)/E(Y) > 1$ holds automatically, with no need for an extra parameter (Xavier et al., 2018). Layering a zero-inflation component onto this base gives the ZIPL, which absorbs excess zeros and overdispersion at the same time. That combination fits group health portfolios well, where most insured individuals claim nothing in a given period, while a few file repeatedly. Xavier et al. (2018) found the ZIPL to outperform the ZIP on real count data across several applications. Against the ZINB, the ZIPL matches the flexibility with one fewer parameter, an edge that matters when claimants are few. All three models, ZIP, ZINB, and ZIPL, are fitted here, with selection guided by AIC, BIC, and goodness-of-fit tests.

Claim severity raises a parallel challenge. Medical claims spring from a mix of routine care and high-cost procedures, and the resulting severity distributions tend to be right-skewed, heavy-tailed, and sometimes multimodal, each shaped by a different category of treatment (Klugman et al., 2019). The pattern is especially sharp in Indonesia, where outpatient visits and inpatient procedures sit far apart in cost (Prastyo & Gani, 2025). Single-component families such as the Exponential, Gamma, Weibull, Lognormal, and Pareto are staples of severity modeling, yet their parametric rigidity makes it hard to fit the body and the tail of heterogeneous claims at once. Miljkovic & Grün (2016) showed empirically that mixture models consistently beat single-component distributions on loss data marked by heterogeneity, heavy tails, and multimodality, precisely because a mixture can build multimodal shapes that map onto distinct claim subpopulations within one framework (Miljkovic & Grün, 2016).

In this study, the Mixture Gamma-Rayleigh Distribution (MGRD) introduced by Bhat & Ahmad (2021) is considered as a candidate distribution for modeling claim severity. The MGRD combines a Gamma component, which can represent lower to moderate claim amounts, with a Rayleigh component that accommodates larger and more dispersed claim amounts. The result is a structure flexible enough to accommodate claims that originate from different healthcare services. Because it folds these distinct claim-generating mechanisms into a single framework, the MGRD should capture heterogeneous medical severity more faithfully than any single-component family. Its shape parameters also let it bend to distributional patterns that a single-component model would miss. Following Omari et al. (2018), the MGRD is benchmarked here against the standard non-negative families, the Exponential, Gamma, Weibull, Lognormal, and Pareto, using goodness-of-fit tests and information criteria to judge how well it suits the observed claims.

Despite the growing interest in aggregate loss modeling, three gaps persist in the literature on Indonesian health insurance pricing. Most studies report only point estimates of the premium and stop short of quantifying its uncertainty. Zero-inflated frequency data have rarely been handled with specialized count distributions such as the ZIPL. And heterogeneous medical severity is still routinely fitted with single-component distributions that overlook the multiple subpopulations within a portfolio.

This study speaks to all three. It applies zero-inflated count models (ZIP, ZINB, and ZIPL) to the full insured population, offers the first empirical test of the ZIPL on Indonesian group health claim frequency data, and builds a confidence interval for the gross renewal premium that folds in both medical inflation and the time value of money.

The interval matters because renewal pricing in group health insurance leans heavily on historical claim experience. A statistically grounded range lets insurers set more competitive premiums tuned to a group's own risk profile, and it gives them room to maneuver at renewal, keeping the premium attractive to insured individuals while still holding enough in reserve against future claim volatility.

METHOD

Research Design

This study uses secondary data from an Indonesian insurance company, specifically from one of its corporate clients participating in a group health insurance product over the period 1 July 2024 to 30 June 2025. Due to confidentiality agreements, the identities of the insurance company and its client are not disclosed. Each insured individual is assigned a unique participant ID to distinguish one from another. The structured data used in this study are presented in Table 1.

Table 1. Data Structure

ID (j)	Claim Arrival (i)	Claim Severity ($X_{j,i}$)	Claim Frequency (N_j)	Total Claim (S_j)
(a)	(b)	(c)	(d)	(e)
A_1	1 2 ⋮ n_1	$x_{1,1}$ $x_{1,2}$ ⋮ x_{1,n_1}	n_1	$\sum_{i=1}^{n_1} x_{1,i}$
A_2	1 2 ⋮ n_2	$x_{2,1}$ $x_{2,2}$ ⋮ x_{2,n_2}	n_2	$\sum_{i=1}^{n_2} x_{2,i}$
⋮	⋮	⋮	⋮	⋮
A_m	1 2 ⋮ n_m	$x_{m,1}$ $x_{m,2}$ ⋮ x_{m,n_m}	n_m	$\sum_{i=1}^{n_m} x_{m,i}$

Based on Table 1, column (a) represents the insured individual number, denoted by j, where $j = 1, 2, \ldots, m$. Each individual filed a number of claims during the coverage period, shown in column (b), representing claim arrivals from the first claim up to the n_j-th claim. The amount paid by the insurance company for each claim is recorded in column (c), denoted by $X_{j,i}$, which represents the claim amount for the i-th claim of the j-th insured individual. Column (d) shows the total number of claims filed by the j-th insured individual, denoted by n_j. Column (e) represents the total claim amount of the j-th insured individual, denoted by S_j.

Population and Sample

The study population consists of 912 insured individuals covered under the group health insurance policy during the observation period. Among these insureds, 102 individuals submitted at least one claim, resulting in a total of 174 observed claims.

Sampling Techniques

Since the data are drawn from a complete administrative record of one corporate client, the entire available claim dataset is used as the study sample. Therefore, no sampling procedure was applied, and all available observations were included in the analysis.

Research Subjects

The study uses secondary data consisting of claim records from the group health insurance policy. The main variables analyzed are: (1) claim frequency, defined as the number of claims submitted by each insured individual during the coverage period, including zero claim counts; and (2) claim severity, defined as the amount paid per individual claim occurrence, measured in Indonesian Rupiah (IDR).

Data Analysis Techniques

The analytical framework in this study follows an aggregate loss modeling approach. The stepwise procedure is as follows.

Aggregate loss represents the total loss that must be borne by an insurance company over a coverage period. Under the collective risk model, aggregate loss S is expressed as the sum of individual claim amounts from a random number of claims N (Klugman et al., 2019):

$$S = X_1 + X_2 + \cdots + X_N; N = 1,2,3, \dots , \quad (1)$$

where S = 0 when N = 0. In this study, the aggregate loss model is developed under the assumptions that $X_1, X_2, \dots, X_N$ are independent and identically distributed claim amounts, and independent of N. The expected aggregate loss per insured is:

$$E[S] = E[X] \times E[N], \quad (2)$$

Under the individual risk model, the aggregate loss is extended from a single insured to a portfolio consisting of m insured units. The total aggregate loss for the portfolio is given by:

$$T = S_1 + S_2 + \cdots + S_m, \quad i = 1,2, \dots , m. \quad (3)$$

Assuming that aggregate losses from different insured units are mutually independent and identically distributed, the expected aggregate loss for the portfolio is given as:

$$E[T] = m \times E[S]. \quad (4)$$

To model aggregate loss, it is necessary to specify distributions for both the number of claims and the size of claims, which will be discussed in the next section. In addition, the parameters of these distributions need to be estimated.

Maximum Likelihood Estimation (MLE) is used to estimate parameters for all distributions considered in this study, as it utilizes the full information in the data and provides efficient estimates (Klugman et al., 2019). Given a random sample $X_1, X_2, \dots, X_n$ of independent and identically distributed observations from a distribution with unknown parameter θ, the likelihood function is defined as:

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta). \quad (5)$$

Where $f(x_i \mid \theta)$ is the probability density function governed by parameter θ. Since the logarithm is a monotonically increasing function, maximizing $L(\theta)$ is equivalent to maximizing its natural logarithm, known as the log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} \ln f(x_i \mid \theta). \quad (6)$$

Claim frequency is the number of claims an insured individual incurs within a coverage period. Being a discrete count, it is naturally modeled with discrete probability distributions, most commonly the Poisson, Binomial, and Negative Binomial (Klugman et al., 2019). However, when the complete insured population is observed, including those who filed no claims, standard distributions may underfit the data due to the presence of excess zeros. In such cases, zero-inflated models are more appropriate, as they account for two distinct sources of zero observations: structural zeros from insured individuals who are inherently unlikely to file claims, and sampling zeros from insured individuals who simply did not file during the period (Qazvini, 2019). This study, therefore, evaluates three zero-inflated candidate models, including ZIP, ZINB, and ZIPL.

The Zero-Inflated Poisson (ZIP) distribution is used to model count data with excess zeros by combining a point mass at zero with a Poisson distribution. For a random variable N following a ZIP distribution with zero-inflation parameter $\pi \in [0,1)$ and Poisson rate parameter $\lambda > 0$, its probability mass function (pmf) is:

$$P(N=n)=\begin{cases}\pi+(1-\pi)e^{-\lambda}, & n=0\\ (1-\pi)\dfrac{\lambda^n e^{-\lambda}}{n!}, & n=1,2,3,\dots\end{cases}. \quad (7)$$

The Zero-Inflated Negative Binomial (ZINB) distribution extends the ZIP distribution by replacing the Poisson component with a Negative Binomial distribution. For a random variable N following a ZINB distribution with zero-inflation parameter $\pi \in [0,1)$, dispersion parameter $r > 0$, and probability parameter $0 < p < 1$, the probability mass function (pmf) is given by:

$$P(N=n)=\begin{cases}\pi+(1-\pi)(p)^r, & n=0\\ (1-\pi)\dfrac{\Gamma(n+r)}{\Gamma(r)n!}p^r(1-p)^n, & n=1,2,3,\dots\end{cases}. \quad (8)$$

The Zero-Inflated Poisson-Lindley (ZIPL) distribution extends the Poisson-Lindley distribution by incorporating a zero-inflation parameter to accommodate excess zeros in count data. The underlying Poisson-Lindley distribution is inherently overdispersed, as its variance exceeds its mean for all parameter values, making the ZIPL distribution suitable for modeling count data exhibiting both excess zeros and overdispersion (Xavier et al., 2018). For a random variable N following a ZIPL distribution with zero-inflation parameter $\pi \in [0,1)$ and shape parameter $\kappa > 0$, the probability mass function (pmf) is given by:

$$P(N=n)=\begin{cases}\pi+(1-\pi)\dfrac{\kappa^2(\kappa+2)}{(\kappa+1)^3}, & n=0\\ (1-\pi)\dfrac{\kappa^2(n+\kappa+2)}{(\kappa+1)^{n+3}}, & n=1,2,3,\dots\end{cases}, \quad (9)$$

Based on the probability mass function in Equation (9), the expected value and variance of the ZIPL distribution are derived as follows:

$$E(N)=(1-\pi)\frac{\kappa+2}{\kappa(\kappa+1)}, \quad (10)$$

$$Var(N)=(1-\pi)\frac{\kappa^3+4\kappa^2+6\kappa+2}{\kappa^2(\kappa+1)^2}+\pi(1-\pi)\left(\frac{\kappa+2}{\kappa(\kappa+1)}\right)^2. \quad (11)$$

All of the model parameters (ZIP, ZINB, ZIPL) were estimated using the maximum likelihood estimation (MLE) method through numerical optimization.

Claim severity is the monetary amount paid on an individual claim. Such data are typically right-skewed. Therefore, non-negative, continuous distributions such as the Exponential, Gamma, Weibull, Pareto, and Lognormal distributions are commonly used for modeling (Klugman et al., 2019; Omari et al., 2018). In addition, this study considers the Mixture Gamma–Rayleigh Distribution (MGRD) proposed by Bhat & Ahmad (2021) to account for potential heterogeneity in medical claim amounts. Each distribution is presented below along with its probability density function and parameter estimation method.

The Exponential distribution is a continuous distribution commonly used to model claim amounts, and is a special case of both the Gamma and Weibull distributions (Omari et al., 2018). A random variable X follows an Exponential distribution with rate parameter $\lambda > 0$ if its probability density function is:

$$f_X(x) = \lambda e^{-\lambda x}, \quad x > 0. \tag{12}$$

The MLE of λ is obtained by setting the derivative of the log-likelihood to zero, yielding:

$$\hat{\lambda} = \frac{1}{\bar{x}} \tag{13}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean (Omari et al., 2018).

The Gamma distribution is a two-parameter extension of the Exponential distribution, expressed in terms of the gamma function $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t}$, $\alpha > 0$ (Omari et al., 2018). A random variable X follows a Gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$ if its probability density function is:

$$f_X(x;\, \alpha, \beta) = \frac{1}{\Gamma(\alpha)\, \beta^{\alpha}}\, x^{\alpha-1}\, e^{-x/\beta}, \tag{14}$$

when $x > 0$.

The MLE of β *is*

$$\hat{\beta} = \bar{x}/\hat{\alpha}, \tag{15}$$

while $\hat{\alpha}$ satisfies the following non-linear equation that is solved numerically:

$$n\left(\ln\hat{\alpha} - \ln\bar{x} - \frac{d}{d\alpha}\ln\Gamma(\hat{\alpha}) + \overline{\ln x}\right) = 0 \tag{16}$$

where $\overline{\ln x} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$ (Omari et al., 2018).

The Weibull distribution is a continuous distribution commonly used to model claim amounts, and reduces to the Exponential distribution when the shape parameter $\gamma = 1$ (Omari et al., 2018). A random variable X follows a Weibull distribution with shape parameter $\gamma > 0$ and scale parameter $\beta > 0$ if its probability density function is:

$$f_X(x;\, \gamma, \beta) = \frac{\gamma}{\beta^{\gamma}} x^{\gamma-1} exp\left[-\left(\frac{x}{\beta}\right)^{\gamma}\right], x > 0 \tag{17}$$

The MLE of γ is obtained by solving:

$$\frac{\sum_{i=1}^{n} x_i^{\hat{\gamma}} \ln x_i}{\sum_{i=1}^{n} x_i^{\hat{\gamma}}} - \frac{1}{\hat{\gamma}} - \frac{1}{n}\sum_{i=1}^{n} \ln x_i = 0. \tag{18}$$

The MLE of β is then:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\hat{\gamma}}\right)^{\frac{1}{\hat{\gamma}}}. \tag{19}$$

A random variable X follows a Lognormal distribution if $\ln X$ is normally distributed. Its probability density function is (Omari et al., 2018).

$$f_X(x;\, \mu, \sigma) = \frac{1}{x\,\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \tag{20}$$

when $x > 0$, where $\mu \in R$ is the location parameter and $\sigma > 0$ is the shape parameter.

The MLEs are:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i\,, \tag{21}$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (\ln x_i - \hat{\mu})^2. \tag{22}$$

The Pareto distribution is a heavy-tailed continuous distribution commonly used to model large claim amounts in insurance applications (Omari et al., 2018). A random variable X follows a Pareto distribution with shape parameter $\alpha > 0$ and scale parameter $x_m > 0$ if its probability density function is:

$$f_X(x;\, \alpha, x_m) = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}, \quad x > x_m. \tag{23}$$

The MLEs of α and x_m are obtained as:

$$\widehat{x_m} = \min\,(x_1, x_2, \dots, x_n) \tag{24}$$

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln\left(\frac{x_i}{\widehat{x_m}}\right)} \tag{25}$$

The Mixture Gamma-Rayleigh Distribution (MGRD) is a two-component mixture of the Gamma and Rayleigh distributions with mixing proportions λ and $1 - \lambda$ respectively (Bhat & Ahmad, 2021). Its flexibility in accommodating varying scale and tail behavior makes it suitable for modeling heterogeneous claim data. Its probability density function is (Bhat & Ahmad, 2021):

$$f_X(x) = \lambda \frac{1}{\beta^{\eta}\Gamma(\kappa)} x^{\eta-1} e^{-x/\beta} + (1-\lambda)\frac{x}{\theta^2} e^{-x^2/(2\theta^2)}, \quad (26)$$

when $\eta, \beta, \theta > 0;\ 0 \leq \lambda \leq 1$, where η and β are the shape and scale parameters of the Gamma distribution, θ is the scale parameter of the Rayleigh distribution, and λ is the mixing proportion.

For the MGRD, the MLE equations are obtained by differentiating the log-likelihood with respect to λ, η, β, and θ (Bhat & Ahmad, 2021):

$$\frac{\partial l}{\partial \lambda} = \sum_{i=1}^{n} \left(\frac{\theta^2 x_i^{\eta-1} e^{-x_i/\beta} - \beta^{\eta}\Gamma(\eta)\, x_i e^{-x_i^2/(2\theta^2)}}{\lambda\theta^2 x_i^{\eta-1} e^{-x_i/\beta} + (1-\lambda)\beta^{\eta}\Gamma(\eta)\, x_i e^{-x_i^2/(2\theta^2)}} \right) = 0, \quad (27)$$

$$\frac{\partial l}{\partial \eta} = \sum_{i=1}^{n} \left(\frac{\lambda\theta^2 x_i^{\eta-1} e^{-x_i/\beta} \begin{bmatrix} \ln(x_i) - \ln(\beta) \\ -\psi(\eta) \end{bmatrix}}{\lambda\theta^2 x_i^{\eta-1} e^{-x_i/\beta} + (1-\lambda)\beta^{\eta}\Gamma(\eta)\, x_i e^{-x_i^2/(2\theta^2)}} \right) = 0, \quad (28)$$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \left(\frac{\lambda\theta^2 x_i^{\eta-1} e^{-x_i/\beta}(x_i - \eta\beta)}{\beta^2 \begin{bmatrix} \lambda\theta^2 x_i^{\eta-1} e^{-x_i/\beta} + \\ (1-\lambda)\beta^{\eta}\Gamma(\eta) \\ x_i e^{-x_i^2/(2\theta^2)} \end{bmatrix}} \right) = 0, \quad (29)$$

$$\frac{\partial l}{\partial \theta} = \sum_{i=1}^{n} \left(\frac{(1-\lambda)\beta^{\eta}\Gamma(\eta)\, x_i e^{-x_i^2/(2\theta^2)}(x_i^2 - 2\theta^2)}{\theta^3 \begin{bmatrix} \lambda\theta^2 x_i^{\eta-1} e^{-x_i/\beta} + \\ (1-\lambda)\beta^{\eta}\Gamma(\eta) \\ x_i e^{-x_i^2/(2\theta^2)} \end{bmatrix}} \right) = 0. \quad (30)$$

Since claim severity represents a monetary amount, it is directly affected by medical inflation. Following the linear scaling property of expectation and variance, if historical claim severity X is scaled by inflation factor (1+I), the inflation-adjusted expected value and variance of the MGRD are given by (Bhat & Ahmad, 2021):

$$E[X] = (1+I)\left(\eta\lambda\beta + (1-\lambda)\theta\sqrt{\frac{\pi}{2}}\right), \quad (31)$$

$$\mathrm{Var}[X] = (1+I)^2 \left(\begin{matrix} \eta(\eta+1)\lambda\beta^2 \\ +2\theta^2(1-\lambda) \\ -\left(\begin{matrix} \lambda\eta\beta + \\ (1-\lambda)\theta\sqrt{\frac{\pi}{2}} \end{matrix} \right)^2 \end{matrix} \right). \quad (32)$$

In this study, I is set to 16.2%, based on the Indonesian medical inflation rate for 2025 as reported by AAJI (2025), ensuring that the premium reflects projected future claim costs rather than historical values. Additionally, a sensitivity analysis is conducted to examine how variations in the inflation rate affect the resulting gross premium, with I evaluated at 10%, 12%, 14%, 16.2%, 18%, and 20%.

To evaluate model adequacy and support model selection, several goodness-of-fit (GoF) tests are applied to the candidate distributions.

For all tests, the hypotheses are defined as H_0: the data follows the specified distribution and H_1: the data does not follow the specified distribution.

For claim frequency, the Chi-Square test is used to compare observed and expected frequencies. The test statistic is given by (Staudte, 2020):

$$\chi^2 = \sum_{i=1}^{d} \frac{(O_i - E_i)^2}{E_i}, \quad (33)$$

where O_i and E_i denote the observed and expected frequencies, respectively.

For claim severity, the Kolmogorov–Smirnov (K–S) and Anderson–Darling (A–D) tests are applied, both of which are based on the cumulative distribution function (cdf). The K-S statistic is (Klugman et al., 2019):

$$D = \max_{t \leq x \leq u} |F(x) - F^*(x)|, \quad (34)$$

where F(x) and F*(x) are the empirical and theoretical cumulative distribution functions, respectively.

The A-D test is an alternative GoF test that is more sensitive to deviations in the tails of the distribution. The test statistic is defined as (Omari et al., 2018):

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F\left(x_{(i)}\right) + \ln\left(1 - F\left(x_{(n-i+1)}\right)\right)\right], \quad (35)$$

where $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ denote the ordered sample, and $F(x)$ is the theoretical cumulative distribution function.

While these tests assess the GoF of each candidate distribution, they do not account for model complexity. To address this, information criteria, namely the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), are also used to compare competing models. These criteria balance model fit and complexity and are defined as (Omari et al., 2018):

$$AIC = -2l + 2k, \quad (36)$$

$$BIC = -2l + k\ln(n), \quad (37)$$

where l is the log-likelihood, k is the number of parameters, and n is the sample size.

In this study, all candidate distributions are evaluated using these criteria. The final model selection is based on a combination of goodness-of-fit results and information criteria, with preference given to models that provide a better balance between fit and parsimony.

To quantify estimation uncertainty, a (1−α) confidence interval for E[X] is constructed via Chebyshev's inequality. This inequality is employed here because it requires no assumptions about the functional form of the underlying distribution, relying solely on the existence of a finite mean and variance (Roos et al., 2022). This is particularly appropriate for insurance claim severity data, which are typically right-skewed and heavy-tailed, rendering normality-based interval methods invalid (Klugman et al., 2019). Although the resulting interval is conservative, this property is desirable in insurance pricing as it guarantees the stated coverage probability regardless of the true distributional form.

Chebyshev's inequality was selected over more commonly applied interval estimation methods for three practical reasons that are specific to the present setting. First, the claim severity data are strongly right-skewed and heavy-tailed, and the fitted MGRD is a two-component mixture; under these conditions, the sampling distribution of the estimated expected severity is not well approximated by a normal distribution, so Wald-type intervals based on asymptotic normality may understate the true uncertainty (Klugman et al., 2019). Second, the number of observed claims is modest (n = 174), which further weakens the large-sample justification required by likelihood-based and normal-approximation intervals. Third, in a renewal-pricing context, a conservative but assumption-light interval is operationally attractive because it provides a coverage guarantee that does not depend on the correctness of the assumed severity distribution, thereby protecting against model misspecification risk during repricing negotiations.

It is nevertheless acknowledged that Chebyshev's inequality is a deliberately conservative choice, and several alternative interval estimation methods could be applied to the same problem, each with distinct advantages and limitations. Bootstrap confidence intervals, particularly the bias-corrected and accelerated (BCa) variant, resample the observed claims directly and can capture the skewness of the severity distribution without imposing a parametric form, often yielding narrower and better-calibrated intervals than distribution-free bounds (Grün & Miljkovic, 2023); however, they rely on the resampled data being representative of the underlying population and may perform poorly in the extreme tail when the number of large claims is small. Likelihood-based intervals, constructed from the profile likelihood or the asymptotic distribution of the maximum likelihood estimator, make efficient use of the fitted MGRD and typically produce tighter intervals, but their validity rests on large-sample asymptotics and on the assumed distribution being correctly specified. Monte Carlo simulation from the fitted aggregate loss model offers the most complete description of premium uncertainty because it can jointly propagate variability from both claim frequency and claim severity and can incorporate parameter uncertainty through a two-stage simulation scheme (Meyers & Schenker, 1983; Venter & Sahasrabuddhe, 2021), but it is computationally more demanding and its output is itself conditional on the estimated parameters and the assumed model structure. Relative to these alternatives, Chebyshev's inequality trades statistical efficiency for robustness: it produces a wider interval, but one whose stated coverage holds under minimal assumptions. The exploration of bootstrap, likelihood-based, and full Monte Carlo intervals is therefore identified as a valuable direction for future refinement of the proposed framework.

$$P\left(\left|E[X]-\widehat{E[X]}\right| \leq \sqrt{\frac{\widehat{Var[X]}}{\alpha n}}\right) \geq 1-\alpha, \qquad (38)$$

giving lower and upper bounds:

$$E[X]_{\mathrm{L}} = \widehat{E[X]} - \sqrt{\frac{\widehat{Var[X]}}{\alpha n}}, \qquad (39)$$

$$E[X]_{\mathrm{U}} = \widehat{E[X]} + \sqrt{\frac{\widehat{Var[X]}}{\alpha n}}. \qquad (40)$$

The pure premium is the expected aggregate claim cost for the coverage period. Under the collective and individual risk frameworks (Equations 2 and 4), and with parameter m representing the number of exposures, the lower and upper bounds of the pure premium are:

$$\varphi_L^P = m\, E[N]\, E[X]_L (1+i)^{-\frac{1}{2}}, \qquad (41)$$
$$\varphi_U^P = m\, E[N]\, E[X]_U (1+i)^{-\frac{1}{2}}, \qquad (42)$$

where i is the insurer's annual interest rate, applied as a discount factor under the assumption that claims are paid, on average, at the midpoint of the coverage period.

Gross premium incorporates a loading factor ω to cover expenses, risk, and profit beyond the expected claim cost:

$$\varphi_L^G = \varphi_L^P \times (1 + \omega), \qquad (43)$$
$$\varphi_U^G = \varphi_U^P \times (1 + \omega). \qquad (44)$$

Figure 1 provides a visual summary of the complete analytical procedure described in the preceding sections, from the input claim data through frequency and severity modeling to the final gross premium estimate.

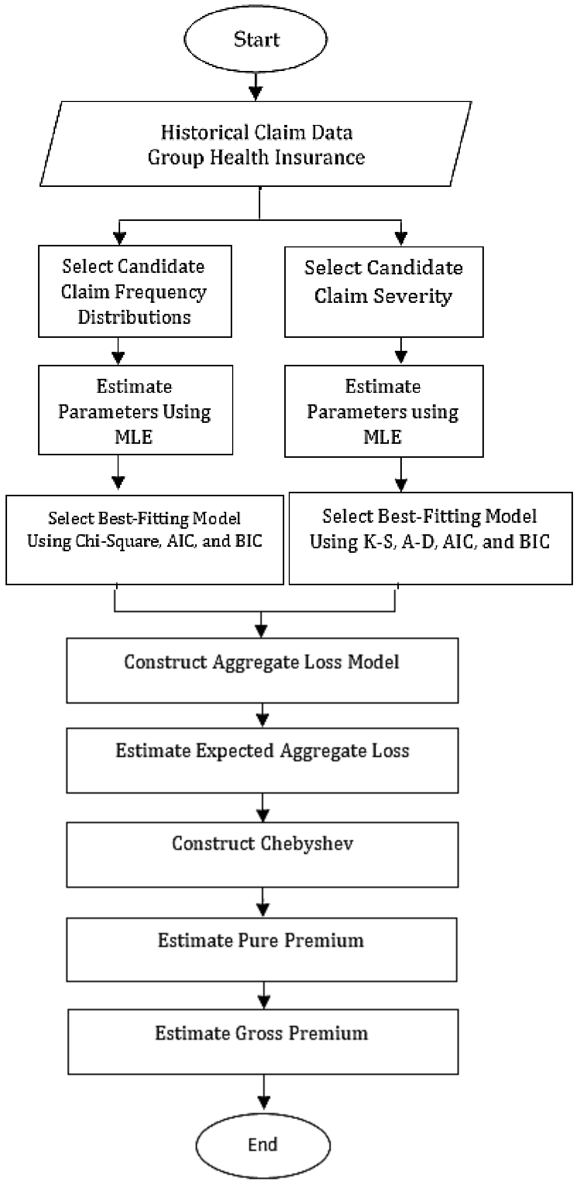

Figure 1. Research Workflow

RESULTS AND DISCUSSION

The data used in this study are the historical claim data of a group health insurance product from one corporate client of an Indonesian insurance company for the period 1 July 2024 to 30 June 2025, as presented in Table 2.

Table 2. Group Health Insurance Claim Data for the Period 1 July 2024 – 30 June 2025 (in IDR)

ID (j)	Claim Arrival (i)	Claim Severity ($X_{j,i}$)	Claim Frequency (N_j)	Total Claim (S_j)
(a)	(b)	(c)	(d)	(e)
A_1	1 2	9,536,626 595,937	2	10,132,563
A_2	1 2 3	22,490,361 16,639,300 31,201,000	3	70,330,661
⋮	⋮	⋮	⋮	⋮
A_{102}	1	11,747,361	1	11,747,361

Based on Table 2, insured individual A_1 filed 2 claims during the period 1 July 2024 to 30 June 2025. The amount paid by the insurance company for the first claim was IDR 9,536,626, and for the second claim was IDR 595,937, bringing the total claim amount for A_1 to IDR 10,132,563. Meanwhile, insured individual A_2 filed 3 claims with a total claim amount of IDR 70,330,661.

Claim Frequency Modeling

Table 3 sets out the claim frequency distribution across the 912 insured individuals. Of these, 102 filed at least one claim during the observation period, and 810 filed none. Among the claimants, 56 filed a single claim, 31 filed two, 8 filed three, 3 filed four, and 4 filed five. The data have a sample mean of 0.1908 and a variance of 0.4026. With 810 of 912 individuals recording no claims, the dataset is heavily zero-dominated, which makes zero-inflated distributions the natural candidates for claim frequency.

Table 3. Claim Frequency Data

Number of Claims	Number of Policyholders
0	810
1	56
2	31
3	8
4	3
5	4
Total	912
Mean	0.1908
Variance	0.4026

To identify the best fit, the ZIP, ZINB, and ZIPL distributions were each fitted to the claim counts in column (d) of Table 2. Fit was assessed with the Chi-Square test at the 5% level, with AIC and BIC as supporting criteria. The parameter estimates and results are presented in Table 4.

Table 4. Claim Frequency Model Comparison

Model	Measure	Value
ZIP	Parameters	$\hat{\pi} = 0.7431$ $\hat{\lambda} = 1.1836$
	Loglik	-438.1376
	AIC	880.2751
	BIC	889.9064
	Chi-Square stat	2.7889
	Chi-Square p-value	0.2480
ZINB	Parameters	$\hat{\pi} = 0.8004$ $\hat{r} = 2.7848$ $\hat{p} = 0.7445$
	LogLik	-437.0521
	AIC	880.1041
	BIC	894.5510
	Chi-Square stat	1.4581
	Chi-Square p-value	0.2272
ZIPL	Parameters	$\hat{\pi} = 0.7431$ $\hat{\kappa} = 1.8221$
	LogLik	-437.3509
	AIC	878.7018
	BIC	888.3331
	Chi-Square	2.0487

Model	Measure	Value
	stat	
	Chi-Square p-value	0.3590

As shown in Table 4, all three candidate models passed the chi-square goodness-of-fit test at the 5% significance level, indicating that each model provides an acceptable representation of the observed claim frequency data. However, the ZIPL model was selected because it achieved the lowest AIC and BIC values while maintaining the highest goodness-of-fit p-value among the competing models. This result suggests that the ZIPL distribution provides the most efficient balance between model fit and complexity for the observed portfolio.

The choice of the ZIPL is well matched to this portfolio, which combines a very large share of zero-claim members (810 of 912) with overdispersion, the variance running well above the mean. The ZIPL absorbs both features at once, which makes it a good fit for group health portfolios where most members never claim. The payoff is a more dependable basis for projecting future claims, exactly what renewal pricing and underwriting decisions rest on.

The diagnostic plots for the fitted ZIPL appear in Figures 2–4. The histogram, the Q-Q plot, and the empirical-versus-theoretical CDF all line up closely with the observed data, reinforcing the goodness-of-fit results in Table 4.

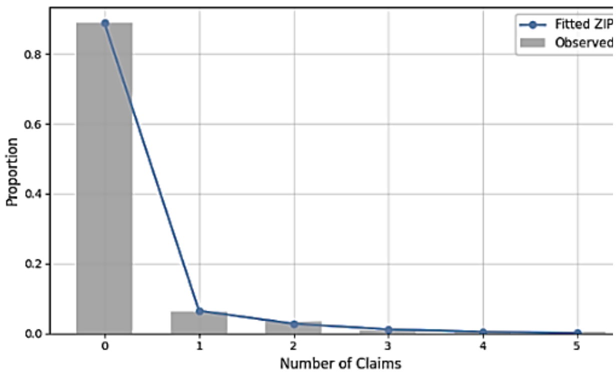

Figure 2. Histogram with Fitted ZIPL

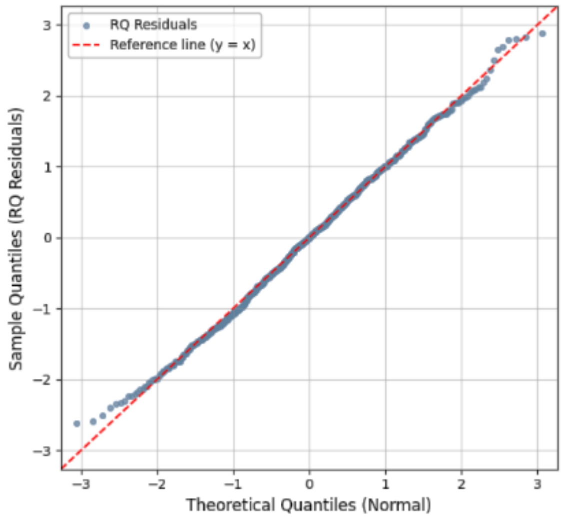

Figure 3. Q-Q Plot of ZIPL

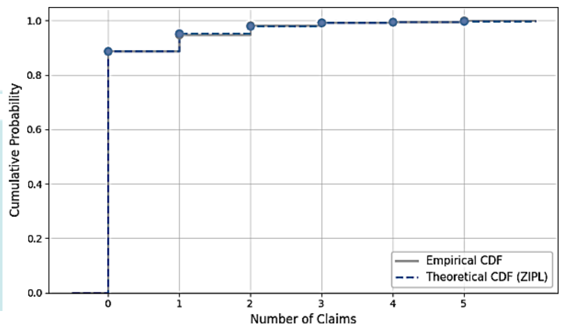

Figure 4. Empirical and Theoretical CDF

Using the estimated parameter of the fitted ZIPL distribution, the expected claim frequency computed using Equation (10) is $\hat{E}[N] = 0.1909$, meaning each insured individual is expected to generate roughly 0.19 claims over the coverage period. Across 912 insured individuals, which works out to about 174 expected claims, squarely in line with what was actually observed, a reassuring sign that the ZIPL captures the portfolio's claim pattern. In practical terms, this expected frequency is one of the two key inputs to the aggregate loss model, since it feeds straight into the expected total claim cost and, with it, into renewal pricing, reserving, and underwriting.

Claim Severity Modeling

The summary statistics for claim severity appear in Table 5. Across 174 observations, the mean claim amount was IDR

13,841,843.52, and the median was IDR 9,449,589.5. The minimum and maximum claim amounts were IDR 150,000 and IDR 61,989,458, respectively, while the standard deviation reached IDR 14,314,917.33. The fact that the mean exceeds the median, together with the wide range of the data and a positive skewness coefficient of 1.3840, confirms that claim severity is positively skewed and widely dispersed. A symmetric model would clearly misrepresent such data, which instead calls for a flexible continuous distribution able to absorb both the skew and the spread.

Table 5. Statistical Summary Claim Severity Data

Total Data	174
Mean	13,841,843.52
Median	9,449,589.5
Minimum	150,000
Maximum	61,989,458
Range	61,839,458
Standard Deviation	14,314,917.33
Skewness	1.3840

To find the best-fitting distribution for claim severity, six distributions were fitted to the claim severity data in column (c) of Table 2, namely Mixture Gamma-Rayleigh (MGRD), Lognormal, Weibull, Gamma, Exponential, and Pareto distributions. Model fit was evaluated using the Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) tests at a 5% significance level, with AIC and BIC used for model comparison. The results are presented in Table 6.

Table 6. Claim Severity Model Comparison

Model	Measure	Value
MGRD	Parameters	$\hat{\lambda} = 0.7672$ $\hat{\eta} = 1.4984$ $\hat{\beta} = 11,890,706.03$ $\hat{\theta} = 595,732.94$
	LogLik	-149.5349
	AIC	307.0698
	BIC	319.7060
	KS stat	0.0304
	KS p-value	0.9956
	AD stat	0.1991
	AD p-value	0.8855
Lognormal	Parameters	$\hat{\sigma} = 1.5388$ $\hat{\mu} = 15.6363$
	LogLik	-181.4888
	AIC	366.9776
	BIC	373.2958
	KS stat	0.1344
	KS p-value	3.34×10^{-3}
	AD stat	5.2551
	AD p-value	5.69×10^{-13}
Weibull	Parameters	$\hat{\gamma} = 0.8341$ $\hat{\beta} = 12,659,985.1742$
	LogLik	-169.3208
	AIC	342.6416
	BIC	348.9598
	KS stat	0.0953
	KS p-value	7.93×10^{-2}
	AD stat	2.5269
	AD p-value	2.23×10^{-6}
Gamma	Parameters	$\hat{\alpha} = 0.7429$ $\hat{\beta} = 18,632,929$
	LogLik	-168.3087
	AIC	340.6174
	BIC	346.9355
	KS stat	0.089432
	KS p-value	0.1163
	AD stat	2.0512
	AD p-value	3.23×10^{-5}
Exponential	Parameters	$\hat{\lambda} \approx 0$
	LogLik	-174.0000
	AIC	350
	BIC	353.1591
	KS stat	0.1521
	KS p-value	5.57×10^{-4}
	AD stat	6.0784
	AD p-value	6.16×10^{-15}
Pareto	Parameters	$\hat{\alpha} = 6.5869$ $\widehat{x_m} = 2,101,422.78$
	LogLik	-200.3014
	AIC	404.6028

Model	Measure	Value
	BIC	410.9209
	KS stat	0.2706
	KS p-value	9.51×10^{-12}
	AD stat	18.0149
	AD p-value	3.70×10^{-24}

As shown in Table 6, the MGRD is the only candidate distribution that passes both the Kolmogorov–Smirnov and Anderson–Darling goodness-of-fit tests at the 5% significance level, with p-values of 0.9956 and 0.8855, respectively. In addition, the MGRD achieves the lowest AIC (307.0698) and BIC (319.7060) among all competing models. The remaining distributions perform less favorably according to both goodness-of-fit statistics and information criteria. Therefore, the MGRD is selected as the most appropriate model for representing claim severity in this portfolio.

The performance of the MGRD is consistent with the heterogeneous nature of medical claim costs, which range from relatively small outpatient claims to substantially larger hospitalization and specialized treatment expenses. Accurate modeling of claim severity is particularly important because severity directly determines the expected aggregate loss and ultimately influences premium determination. Consequently, the selected MGRD provides a more reliable basis for renewal pricing, reserve allocation, and risk management. By better capturing the variability of medical claim amounts, the model helps insurers anticipate potential exposure to high-cost claims and supports more informed underwriting decisions.

Figures 5–7 present the diagnostic plots for the fitted MGRD model. The fitted density, Q-Q plot, and empirical versus theoretical CDF show close agreement between the observed data and the fitted distribution. These results are consistent with the goodness-of-fit statistics in Table 6 and further support the suitability of the MGRD for modeling claim severity.

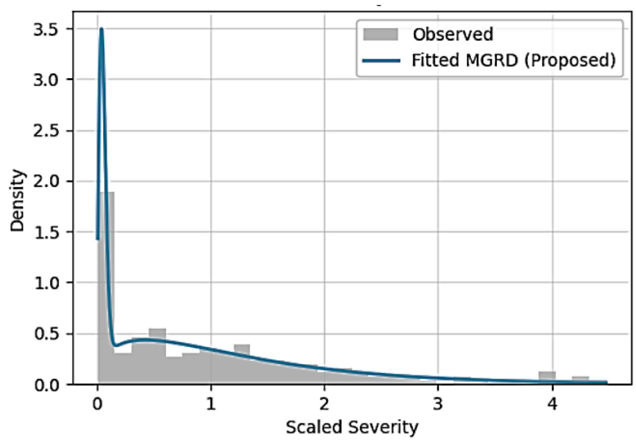

Figure 5. Histogram with Fitted MGRD

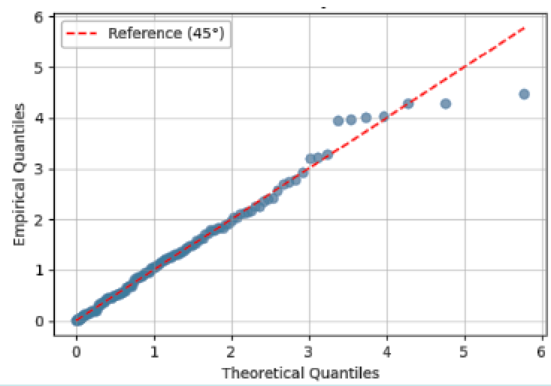

Figure 6. Q-Q Plot of MGRD

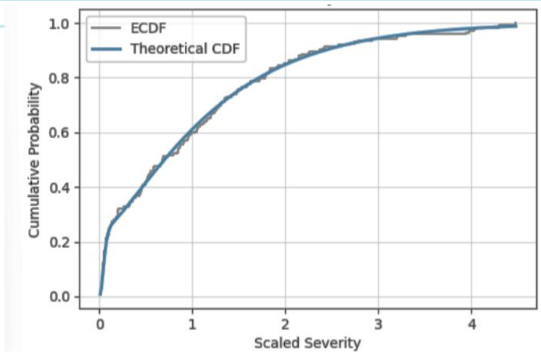

Figure 7. Empirical and Theoretical CDF of MGRD

Using the fitted MGRD and assuming a medical inflation rate of 16.2%, the inflation-adjusted expected claim severity from Equation (31) is estimated at $\widehat{E[X]} =$ 16,086,583.77, and the variance from Equation (32) is $\widehat{Var[X]} =$ 289,790,308,046,754.69. From a risk management standpoint, this implies a need to hold adequate reserves and to incorporate the uncertainty into renewal pricing. An insurer that disregards it remains exposed to unexpected claim deterioration, particularly where exposure to hospitalization and other high-cost treatment is substantial.

To quantify this uncertainty, a 95% confidence interval for the expected claim severity was then constructed using Chebyshev's inequality (Equations 39–40) at α = 5% with n = 174 claims:

$$E[X]_{\mathrm{L}} = \widehat{E[X]} - \sqrt{\frac{\widehat{Var[X]}}{\alpha n}}$$
$$= 16{,}086{,}583.77 - \sqrt{\frac{289{,}790{,}308{,}046{,}754.69}{(0.05)(174)}}$$
$$= IDR\ 10{,}315{,}168.80.$$

$$E[X]_{\mathrm{U}} = \widehat{E[X]} + \sqrt{\frac{\widehat{Var[X]}}{\alpha n}}$$
$$= 16{,}086{,}583.77 + \sqrt{\frac{289{,}790{,}308{,}046{,}754.69}{(0.05)(174)}}$$
$$= IDR\ 21{,}857{,}998.75.$$

It is important to be explicit about which sources of uncertainty are captured by the resulting premium interval. In an aggregate loss framework, total losses are driven by both claim frequency and claim severity, and the uncertainty in each component can be decomposed into process risk, which is the inherent randomness of the loss-generating process, and parameter risk, which arises because the model parameters are estimated from a finite sample (Meyers & Schenker, 1983; Venter & Sahasrabuddhe, 2021). In the present study, the premium interval is derived from the Chebyshev bound on the expected claim severity and therefore reflects the dispersion (process variability) of the severity distribution around its mean. Claim frequency enters the premium calculation only through its point estimate of the expected number of claims, E[N], so the interval does not propagate the additional variability associated with the frequency component. Likewise, because the interval is built from the fitted MGRD using its estimated parameters, parameter risk in both the frequency and severity models is not separately quantified; the estimated parameters are treated as fixed inputs. Consequently, the reported interval should be interpreted primarily as a severity-driven, process-oriented range rather than a complete characterization of aggregate loss uncertainty. This is a deliberate simplification adopted for tractability and transparency, and it is likely to understate the full uncertainty of the renewal premium. A complete decomposition of process and parameter risk across both the frequency and severity components, for example, through a two-stage Monte Carlo simulation of the compound model, is identified as an important extension and is revisited in the limitations and future research discussion.

Pure Premium Estimation

The fitted ZIPL–MGRD compound model is used to estimate the pure premium for the renewal period. With the insurer's annual interest rate of 6.5% and m = 912 insured, the lower and upper bounds of the pure premium from Equations (41)–(42) are:

$$\varphi_L^P = m\,\widehat{E[N]}\,\widehat{E[X]}_L(1+i)^{-\frac{1}{2}}$$
$$= 912 \times 0.1909 \times 10{,}315{,}168.80 \times (1+6.5\%)^{-\frac{1}{2}}$$
$$= IDR\ \ 1{,}740{,}441{,}901.56.$$
$$\varphi_U^P = m\,\widehat{E[N]}\,\widehat{E[X]}_U(1+i)^{-\frac{1}{2}}$$
$$= 912 \times 0.1909 \times 21{,}857{,}998.75 \times (1+6.5\%)^{-\frac{1}{2}}$$
$$= IDR\ 3{,}688{,}022{,}721.40.$$

Gross Premium Estimation

The gross premium is obtained by applying a loading factor to the estimated pure premium bounds. The loading structure follows the insurer's internal pricing practice and consists of profit margin, operating expenses, and sales commission, as shown in Table 7.

Table 7. Loading Component

Loading Component	Percentage (%)
Profit Margin (ω_1)	5.00
Operating Expenses (ω_2)	8.00
Sales Commission (ω_3)	5.00
Total (ω)	18.00

Applying the 18% loading factor from Table 7 using Equations (43)–(44):

$$\begin{aligned}\varphi_L^G &= \varphi_L^P \times (1 + \omega)\\ &= 1{,}740{,}441{,}901.56 \times (1 + 18\%)\\ &= IDR\ 2{,}053{,}721{,}443.84.\\ \varphi_U^G &= \varphi_U^P \times (1 + \omega)\\ &= 3{,}688{,}022{,}721.40 \times (1 + 18\%)\\ &= IDR\ 4{,}351{,}866{,}811.26.\end{aligned}$$

For the renewal coverage period, the estimated gross premium ranges from IDR 2,053,721,443.84 to IDR 4,351,866,811.26. This interval is higher than the gross premium charged during the previous policy period (IDR 2,439,280,800), during which total claims reached IDR 2,408,480,772, resulting in a loss ratio of approximately 98.74%. This indicates that nearly all premium income was used to cover claim payments, leaving limited margins for operating expenses, commissions, and profit. The estimated premium interval, therefore, suggests that a higher renewal premium may be required to maintain pricing adequacy. The relatively wide interval reflects the uncertainty surrounding future claim costs, particularly the substantial variability in medical claim severity. In practice, this range can support renewal negotiations by providing actuarially justifiable premium options that balance competitiveness with financial sustainability.

The proposed interval can also be set against the pricing approach the insurer currently applies in practice. In its renewal process, the company prices premiums deterministically, taking the total claims incurred in the previous period as the basis for projecting the claims expected in the coming period and adjusting this figure for medical inflation together with an incurred-but-not-reported (IBNR) loading to allow for claims that have not yet been reported. Applied to the present portfolio, this deterministic procedure produces a single renewal premium of IDR 3,573,076,486.24. Because it yields only one figure, the deterministic approach cannot convey the uncertainty surrounding total claims; yet in insurance, the quantity that matters is not only the expected level of claims but also the chance that total claims turn out to be far higher than expected. The aggregate loss distribution speaks to this directly, describing the probabilities attached to a range of possible total-claim outcomes within a period, much as the severity distribution describes the probabilities attached to the size of an individual claim (Klugman et al., 2019). The deterministic premium falls within the estimated interval, positioned closer to the upper end, a result consistent with the conservative inflation and IBNR loadings built into the company's method. That the deterministic figure sits comfortably inside the proposed range lends external support to the estimated interval, while the interval itself has the further merit of expressing premium adequacy as a probabilistic range rather than a single point.

The validation of the estimated premium in this study is based on two complementary benchmarks: a comparison of the proposed premium interval with the gross premium charged in the immediately preceding policy period, together with the realized loss ratio of that period, and a comparison with the deterministic premium produced by the insurer's own traditional pricing method, as discussed above. These benchmarks provide a meaningful initial check of pricing adequacy because they show whether the proposed range would have been sufficient to cover the claims actually incurred and how it relates to the premium the company would have charged using its established approach. It is acknowledged, however, that comparison against a single historical period and a single alternative method is still a limited form of validation and does not, on its own, establish the predictive performance of the framework. More rigorous validation procedures, such as out-of-sample validation in which the model is calibrated on one part of the experience and tested on another, and retrospective testing across several historical policy periods, would provide stronger evidence of predictive accuracy (Frees, 2018). These procedures could not be carried out in the present study because only a single year of claim experience for one corporate client was available, which does not permit a temporal train-and-test split or multi-period backtesting. These additional validation analyses are therefore recommended as a priority for future research once longer claim histories and data from multiple policy periods or clients, become available.

Table 8 reports how the premium responds to medical inflation assumptions between 10% and 20%, against the 16.2% baseline. The direction is unsurprising: higher inflation lifts expected severity and, with it, the renewal premium. The gross interval widens from IDR 1.944–4.120 billion at 10% to IDR 2.121–4.494 billion at 20%. What stands out is the stability of the pattern: although the numbers shift across scenarios, the steady upward march suggests the pricing model is fairly robust to the inflation assumption. The practical lesson is to keep a close eye on medical inflation, since sustained cost growth can move premium requirements and portfolio profitability, in a hurry.

It is worth clarifying how medical inflation enters the model. In this study, it is applied as a scaling factor on claim severity, so the sensitivity analysis adjusts the expected size of each claim and, through it, the estimated premium, while the claim frequency is held at its estimated value. This treatment follows the way medical claim costs are conventionally decomposed, in which total benefits paid can be expressed as the average cost per claim multiplied by the number of claims, so that the cost-per-service or unit-price component, which is precisely what medical inflation measures, operates on the severity dimension rather than on the count of claims (Klugman et al., 2019). Modeling inflation on severity is therefore consistent with how rising healthcare prices propagate into claim amounts, and it is in line with how medical inflation is characterized in the Indonesian health insurance setting as growth in the cost of medical services (Prastyo & Gani, 2025). Restricting inflation to the severity channel also keeps the sensitivity analysis transparent, since the resulting variation in the premium can be attributed unambiguously to changes in claim cost rather than to a confounded mix of cost and utilization effects.

At the same time, it should be acknowledged that medical inflation may also influence claim frequency indirectly. Rising healthcare costs, the expansion of available treatments, and shifts in care-seeking behavior can change how often insured individuals utilize medical services, and benefit redesign undertaken in response to cost pressure can likewise alter utilization patterns. The present framework does not capture this potential feedback of inflation on frequency, so the sensitivity results should be read as reflecting the severity channel of medical inflation only. Incorporating utilization-driven frequency effects would provide a more complete picture of how medical inflation propagates into renewal premiums and is identified as a useful direction for future work.

Table 8. Medical Inflation Sensitivity Analysis

Medical Inflation	Expected Severity (Million IDR)	Gross Premium Lower (Billion IDR)	Gross Premium Upper (Billion IDR)
10%	15.23	1.944	4.120
12%	15.51	1.979	4.195
14%	15.78	2.015	4.269
16.2%	16.09	2.053	4.352
18%	16.34	2.086	4.419
20%	16.61	2.121	4.494

CONCLUSION

Conclusion

This study applies an aggregate loss modeling approach to estimate the renewal premium for a group health insurance portfolio in Indonesia using historical claim data from July 2024 to June 2025. Claim frequency is modeled by the Zero-Inflated Poisson–Lindley distribution with $\hat{\pi} = 0.7431$ and $\hat{\kappa} = 1.8221$, selected based on the highest chi-square p-value and the lowest AIC and BIC among the candidate models. Claim severity is modeled by the Mixture Gamma–Rayleigh Distribution with $\hat{\lambda} = 0.7672$, $\hat{\eta} = 1.4984$, $\hat{\beta} = 11{,}890{,}706.03$, and $\theta = 595{,}732.94$, as it was the only model to pass both the K-S and A-D goodness-of-fit tests while achieving the lowest AIC and BIC. Under a 95% confidence level and a medical inflation assumption of 16.2%, the estimated renewal gross premium ranges from IDR 2.053 billion to IDR 4.352 billion. The width of this interval is a feature rather than a limitation, as it honestly communicates the risk range insurers face when pricing renewal premiums. From a practical perspective, the estimated premium interval provides insurers with an actuarially justified range that can support renewal pricing and premium negotiations with policyholders. Rather than relying on a single premium estimate, insurers may use the interval to balance pricing adequacy, financial sustainability, and market competitiveness. The results also highlight the importance of considering claim severity uncertainty and medical inflation when determining renewal premiums, particularly in portfolios exposed to high-cost medical claims.

Sugestion

Future research may extend this study by incorporating Bayesian credibility models to combine portfolio-specific experience with broader insurance data. Dependence structures between claims or between frequency and severity may also be explored to better capture claim dynamics. In addition, dynamic medical inflation models could be considered to reflect changing healthcare cost trends over time. Finally, the proposed framework may be validated using larger datasets obtained from multiple insurance companies and policyholder groups to enhance the robustness and generalizability of the results.

BIBLIOGRAPHY

Asosiasi Asuransi Jiwa Indonesia (AAJI). (2025). AAJI Daily News - 31 Juli 2025. AAJI. https://aaji.or.id/Berita/aaji-daily-news---31-juli-2025

Bhat, A. A., & Ahmad, S. P. (2021). Mixture of Gamma and Rayleigh distributions: Mathematical properties and applications. *Journal of Applied Probability and Statistics*, 16(2), 81–97. https://www.researchgate.net/publication/354047993_Mixture_of_Gamma_and_Rayleigh_Distributions_Mathematical_Properties_and_Applications

Frees, E. W. (Ed.). (2018). Loss data analytics: An open text authored by the actuarial community. https://openacttexts.github.io/Loss-Data-Analytics/

Grün, B., & Miljkovic, T. (2022). The automated Bias-Corrected and accelerated Bootstrap confidence intervals for risk measures. *North American Actuarial Journal*, 27(4), 731–750. https://doi.org/10.1080/10920277.2022.2141781

Khandeka, G., & Shaikh, A. (2022). A Study on Medical Inflation Affects Healthcare Costs and Quality. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, 2(5). 189-193. https://ijarsct.co.in/Paper14654.pdf

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2019). Loss Models: From Data to

Decisions (5th ed.). John Wiley & Sons.

Kusumadewi, R., Riaman, R., & Sukono, S. (2022). Determining the price of fisherman micro insurance premiums using the Aggregate Risk Model approach in Cirebon Regency. *International Journal of Quantitative Research and Modeling*, 3(3), 118–123. https://doi.org/10.46336/ijqrm.v3i3.346

Lestari, P. A. I. (2022). Simulasi Monte Carlo untuk prediksi jumlah klaim asuransi di BPJS Ketenagakerjaan Cabang Bojonegoro. Jurnal Statistika Dan Komputasi, 1(2), 93–100. https://doi.org/10.32665/statkom.v1i2.1265

Manurung, T. (2025). Distribution Models of Claim Frequency and Claim Severity in Determining the Pure Premium of Car Insurance with the Application of a Deductible. *ZERO Jurnal Sains Matematika Dan Terapan*, 9(2), 662. https://doi.org/10.30829/zero.v9i2.26257

Meyers, G., & Schenker, N. (1983). Parameter uncertainty in the collective risk model. Proceedings of the Casualty Actuarial Society, 70, 111–143. https://www.casact.org/sites/default/files/database/proceed_proceed83_83111.pdf

Miljkovic, T., & Grün, B. (2016). Modeling loss data using mixtures of distributions. *Insurance Mathematics and Economics*, 70, 387–396. https://doi.org/10.1016/j.insmatheco.2016.06.019

Oh, R., Shi, P., & Ahn, J. Y. (2019). Bonus-Malus premiums under the dependent frequency-severity modeling. *Scandinavian Actuarial Journal*, 2020(3), 172–195. https://doi.org/10.1080/03461238.2019.1655477

Omari, C. O., Nyambura, S. G., & Mwangi, J. M. W. (2018). Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions. *Journal of Mathematical Finance*, 08(01), 137–160. https://doi.org/10.4236/jmf.2018.81012

Prastyo, C. E., & Gani, A. (2025). Medical cost inflation and its drivers in Indonesian employer-sponsored health insurance for retiree families. *Narra J*, 5(2), e2528. https://doi.org/10.52225/narra.v5i2.2528

Qazvini, M. (2019b). On the Validation of Claims with Excess Zeros in Liability *Insurance: A Comparative Study. Risks*, 7(3), 71. https://doi.org/10.3390/risks7030071

Roos, E., Brekelmans, R., Van Eekelen, W., Hertog, D. D., & Van Leeuwaarden, J. S. (2021). Tight tail probability bounds for distribution-free decision making. *European Journal of Operational Research*, 299(3), 931–944. https://doi.org/10.1016/j.ejor.2021.12.010

Selvakumar, V., Satpathi, D. K., Kumar, P. T. V. P., & Haragopal, V. V. (2023). Modeling of loss severity of motor insurance extreme claims. *International Journal of Advances in Soft Computing and Its Applications*, 15(1). https://www.researchgate.net/publication/370292197_Modeling_of_Loss_Severity_of_Motor_Insurance_Extreme_Claims

Staudte, R. G. (2020). Evidence for goodness of fit in Karl Pearson chi-squared statistics. *Statistics*, 54(6), 1287–1310. https://doi.org/10.1080/02331888.2020.1862115

Venter, G. G., & Sahasrabuddhe, R. (2021). A note on parameter risk. *Variance*, 9(1), 54–63. https://www.casact.org/sites/default/files/2021-07/Parameter-Risk-Venter-Sahasrabuddhe.pdf

Xavier, D., Santos-Neto, M., Bourguignon, M., & Tomazella, V. (2017). Zero-Modified

Poisson-Lindley distribution with applications in zero-inflated and zero-deflated count data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1712.04088