mathematics

MDPI

Article

Frequency–Severity Dependence Diagnostics in Motor Casco Insurance: From Measurement Bias to Risk-Based Premium Pricing and Claims Reserving Strategies

Yusuf Arslan [1,2,*] and Filiz Karaman [1]

[1] Department of Statistics, Yıldız Technical University, 34220 Istanbul, Türkiye; fkaraman@yildiz.edu.tr
[2] Actuarial Department, Türk Reasürans, 34782 Istanbul, Türkiye
* Correspondence: yusuf.arslan1@std.yildiz.edu.tr

Abstract

Diagnostics for frequency–severity dependence in insurance can indicate a positive association even when claim counts and claim sizes are independent, because the association is generated by how the claims are summarized rather than by the claims process. We study the association between a policy's claim count and its average claim size, which arises by construction, give its mechanism on the rank scale, and decompose it in a Turkish motor casco portfolio of 818,789 policies and 214,283 claims: the observed count–average Kendall's τ of +0.08 is reproduced in full by permuting severities across claims (+0.11 under exact independence), so that a copula fitted to this pair would absorb an association that does not belong to the loss process. We then propose a leakage-safe diagnostic protocol that combines claim-level rank measures, covariate-adjusted regression, cluster-robust inference with a power analysis, an equivalence test, and a replicated synthetic positive control. Clean measurement leaves a small negative association ($\tau = -0.0065$) whose effect on a Solvency-II-style capital proxy cannot be distinguished from zero, whereas injected dependence is reliably detected. In the decision layer, modeling the pure premium rather than frequency alone drives risk differentiation, a trend-based on-level adjustment restores the premium level, and the paid, incurred, and case-reserve bases of the book are reconciled before a reserve confidence level is chosen. The practical message is that measurement should precede model complexity: a dependence layer is warranted only after it survives artifact-safe measurement.

Keywords: motor insurance; frequency–severity dependence; measurement artifacts; copula diagnostics; Monte Carlo simulation; premium pricing; claims reserving

MSC: 62P05; 91G05; 62H05; 62G32

Check for updates

Academic Editors: Linyi Qian and Zhuo Jin

Received: 28 August 2026
Revised: 27 October 2026
Accepted: 2 November 2026
Published: 9 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Classical collective-risk models write a policy's total claim cost as $S = Y_1 + \cdots + Y_N$, with the claim count N independent of the claim costs Y_i. Much actuarial work, including longitudinal claims models and experience rating [1,2], relaxes this assumption, mostly through copulas [3–6], which need care with discrete counts [7]. In motor insurance, reported signs vary: negative count effects on severity in collision and comprehensive coverages [8,9], positive associations for liability claims in dependent two-part models [10] and for aggregated count–average-size pairs [11,12]. Ignoring genuine dependence biases the risk

premium [2,9] and can misstate the capital-relevant aggregate-loss tail [13] and reserve uncertainty, as documented for misjudged dependence between lines of business [14].

Dependence is a property of the joint distribution of counts and severities, association a sample statistic of observed quantities, such as Kendall's τ. An association is spurious if it is generated by how the quantities are constructed rather than by the claims process. Two constructions concern us. Because S has N summands, the count and the sum are positively associated even under independence whenever severities are nonnegative with positive mean, and the count is non-degenerate; this count–sum property is elementary and is stated only for completeness (Text S1, Section S1). The average claim size S/N also shares N: its conditional mean is constant in N, but for right-skewed severities its median typically rises with N, and rank measures respond (aggregation-induced rank association; Remark 1). Oh et al. [15] show that the count–average pair of copula models for summarized claims [11,12] is dependent even when frequency and individual severities are independent, because the variance of the average scales with $1/N$ and propose models that accommodate this structure; bonus hunger alone can likewise induce dependence in reported claims [16]. We do not claim that joint modeling of the count–average pair is an error. What has not been established is the sign and size of the construction-induced association on the rank scale, and how much of an observed count–average association in a real portfolio it accounts for: if a copula fitted to that pair absorbs an association that exact independence reproduces, its sign and strength are not evidence about frequency–severity dependence and cannot be carried into a claim-level simulator. A third construction, the reuse of the count's uniform variate to select severity quantiles in a compound simulator, is an implementation error rather than a measurement problem; it is treated as a safeguard of the simulation protocol (Text S1, Section S1), where its consequences are quantified.

Against this background, three gaps remain. First, the literature does not separate dependence from association induced by how claims are summarized, so reported signs cannot be compared on a common footing. Second, no validated, leakage-safe measurement protocol exists whose null result is informative. Third, the economic consequence of a small dependence for capital is rarely quantified. Our research question is whether apparent frequency–severity dependence survives artifact-safe measurement and, if any remains, whether it materially affects aggregate capital in a short-tail, attritional motor own-damage (casco) book; pricing and reserving implications are secondary. The protocol is line-independent; the answer need not transfer to long-tail or catastrophe-exposed lines (Section 5.3). We address these gaps with three contributions. (i) We give the mechanism of the count–average association on the rank scale, measure it under a known truth, and decompose the observed association of a real portfolio into the part reproduced under exact independence and the residual, which identifies the construction that cannot answer the research question. (ii) We apply a leakage-safe protocol (Algorithm 1) to the portfolio, separating the pricing-relevant frequency–severity axis from the reserving-relevant settlement-lag axis, with cluster-robust inference, a power analysis, an equivalence test, and a replicated synthetic positive control, so that a small or null estimate is informative and reflects the data rather than the method. (iii) A paired, leakage-free capital experiment measures the capital effect of what remains and the association at which that effect would become material. A secondary decision layer, a 2×2 pricing comparison with a return on risk-adjusted capital (RORAC) premium target and a reconciliation of reserve bases, shows where modeling effort pays off once the dependence question is settled.

Empirically, artifact-safe measurement turns a strong positive naive association into a small negative one whose capital effect is immaterial for the portfolio examined; for this attritional motor portfolio, modeling effort is better spent on risk differentiation than on a dependence margin. Section 2 presents the data and methods, Section 3 the results,

Section 4 the discussion, and Section 5 the conclusions, recommendations for practice, and limitations. Appendix A reports the full covariate-adjusted regression and Appendix B the specification of every pricing model; formal statements, proofs, implementation details, the premium-principle and reserve-strategy material, and an independent re-derivation of the headline results are collected in Supplementary Text S1.

Algorithm 1. Leakage-safe frequency–severity diagnostic protocol

Input: policy records, claim-level severities, exposures, covariates, CPI.

1. Deflate claim payments and case reserves to a common price basis.
2. Measure dependence only on the clean axes, count × per-claim severity and settlement delay × severity. A count × total or count × average pair may be modeled jointly, but its association is not by itself evidence of dependence between frequency and individual severity: decompose any such naive measure against a permutation null before interpreting it (Section 3.2).
3. Estimate the covariate-adjusted count effect; test against independence with policy-cluster bootstrap intervals, and report the cluster-adjusted minimum detectable effect and an equivalence test (Section 2.4).
4. Validate on synthetic portfolios with injected τ: require nominal Type-I error, monotone recovery, and, with the claim-severity margin fixed, a material capital response at strong injection but none at zero injection.
5. Simulate capital without reusing the count uniform (Text S1, Proposition S2), through a copula calibrated to the observed claim-level association with the claim-severity margin fixed; carry dependence into premium and reserve rules only if it changes the portfolio capital distance K (99.5% quantile minus mean) by at least 5% (Section 2.7).

2. Materials and Methods

After the data and workflow, Section 2.3 gives the mechanism of the construction-induced association and the controlled experiment that quantifies it; Sections 2.4 and 2.5 define and validate clean measurements; Sections 2.6 and 2.7 translate dependence into capital, and Sections 2.8–2.10 cover the secondary decision layer (pricing, the premium target, and the reconciliation of reserve bases). Algorithm 1 summarizes the protocol.

2.1. Data and Constant-Price Transformation

2.1.1. Data Files, Linking, and Cleaning

The data are three administrative files of the motor own-damage (casco) portfolio of a Turkish insurer: a policy file with 818,789 contracts underwritten in 2014–2018, a paid-claims ledger with 432,570 individual payments, and a case-reserve ledger with 135,758 quarter-end snapshots. Payments and reserves are observed up to December 2020, so even the youngest underwriting year has at least two years of run-off. Accident dates extend into 2019, because policies written toward the end of 2018 remain in force during the following calendar year; the paid triangle therefore covers six accident years, 2014–2019. The rating variables are vehicle age, insured age, horsepower, engine displacement, insured vehicle value, gender, and body type, recorded together with earned exposure and the collected casco premium.

The preparation is fully scripted. Records are typed and validated at ingestion; policies with missing underwriting dates or with underwriting years outside 2014–2018 are removed; claims are identified by unique claim-file numbers and linked to policies through the policy reference, so that multiple payment rows represent development rather than

duplication and are aggregated per claim file; payment records with missing dates are dropped, development years are constrained to be nonnegative, and claim files with nonpositive cumulative real payment are excluded from the severity analyses. In the present extract, these rules removed no policy, payment, or reserve record (818,789 policies, 432,570 payment rows, and 135,758 snapshots before and after cleaning). Every claim file links to exactly one policy and one accident year, and the covariate fields have no missing values. Linking claims to policies yields 214,568 claim files, all counted in the policy claim count N; the only exclusion concerns the 285 files (0.13%; 30 with zero and 255 with negative cumulative real payment), which are dropped from the severity analyses only, leaving a severity sample of 214,283 claims.

Negative payment entries (65,724 of the 432,570 rows) represent salvage and subrogation recoveries. They are kept because removing them would distort both the ultimate losses and the development pattern. No claim amount is capped, trimmed, or winsorized: unusually large claims are retained and modeled explicitly by the generalized Pareto tail (Section 2.6), and for pricing, policy-level ultimate costs that net to a negative value after recoveries are floored at zero (Section 2.8).

2.1.2. Constant-Price Transformation

Given high, variable inflation in Türkiye, claim payments and case reserves are expressed in Turkish lira (TRY) at December 2018 prices using the monthly consumer price index (CPI). Collected premiums are deflated separately and are used only for the relative adequacy diagnostic (Section 2.9; Text S1, Figure S1); because the scope of the recorded premium field cannot be independently verified, they are not summarized in Table 1. The insured vehicle value is recorded at underwriting-year (nominal) prices and enters the analyses as a nominal rating variable without deflation. As a sensitivity check, shifting the CPI slope by −5% and +5% changes real incurred losses (paid amounts plus case reserves open at the valuation date) by −7.0% and +7.8%, respectively, and leaves the qualitative conclusions unchanged. Deflation also leaves the shape of the severity distribution essentially unchanged (Table 1C).

Table 1. Descriptive statistics of the portfolio. Panel (**A**): policy-level rating covariates (818,789 policies); P5 and P95, 5th and 95th percentiles; insured age zero almost exclusively identifies corporate policyholders. Panel (**B**): composition by underwriting year; claims are all linked claim files (including the 285 excluded from the severity sample); claim frequency is claims per policy-year of earned exposure. Panel (**C**): claim severity of the severity sample (214,283 claims) at nominal and December-2018 prices. CV, coefficient of variation; P90–P99, percentiles; SD, standard deviation; TRY, Turkish lira.

Panel (A). Rating Covariates					
Variable	**Mean**	**SD**	**P5**	**Median**	**P95**
Vehicle age (years)	3.32	2.95	0	3	9
Insured age (years)	36.8	20.7	0	39	67
Horsepower	115.0	34.3	75	110	177
Engine displacement (cm^3)	1545	297	1198	1560	1995
Insured vehicle value (TRY, nominal)	76,656	64,593	24,240	60,550	183,439
Earned exposure (years)	0.99	0.05	0.99	1.00	1.00

Table 1. *Cont.*

Panel (B). Composition by Underwriting Year					
Underwriting Year	**Policies**	**Earned Exposure (Policy-Years)**	**Claims**	**Claim Frequency**	**Median Vehicle Value (TRY, Nominal)**
2014	160,545	160,544	47,156	0.294	45,000
2015	196,301	196,300	54,044	0.275	54,055
2016	166,090	166,086	42,473	0.256	60,900
2017	160,928	160,902	40,332	0.251	68,035
2018	134,925	127,118	30,563	0.240	81,630
Total	818,789	810,950	214,568	0.265	60,550

Panel (C). Claim Severity (TRY)							
Price Basis	**Mean**	**SD**	**Median**	**P90**	**P95**	**P99**	**CV**
Nominal	3297	8091	1134	7616	13,447	34,449	2.45
Real (December 2018)	4243	9895	1529	9870	17,220	43,152	2.33

2.1.3. Portfolio Composition and Claim Severity

Table 1 summarizes the numeric rating covariates (Panel A), the composition of the portfolio by underwriting year (Panel B), and claim severity at both price bases (Panel C). Earned exposure is close to one year for almost all policies. Claim frequency per unit of exposure declines steadily from the oldest to the youngest underwriting year, from 0.294 in 2014 to 0.240 in 2018, while the median nominal vehicle value rises from TRY 45,000 to TRY 81,630. Policyholders are 59.1% male, 23.7% female, and 17.1% corporate; corporate policyholders are, with few exceptions, recorded with insured age zero, and virtually all policies (all but fewer than ten) share one body type, so the body-type field carries no effective rating information (Appendix A). Of all policies, 79.0% have no claim, 16.8% one, 3.3% two, and 0.8% three or more (mean 0.262, variance-to-mean ratio 1.23). By accident year, 2014–2019 contribute 21,416; 54,316; 46,298; 41,312; 36,287; and 14,654 claims in the severity sample, respectively.

At both price bases, claim severity is heavily right-skewed, with the mean far above the median and a long upper tail: in real terms, the mean is TRY 4243 against a median of TRY 1529 and a 99th percentile of TRY 43,152. Deflation rescales the distribution without changing its shape materially, as the two coefficients of variation show (2.33 real against 2.45 nominal).

2.1.4. Paid Triangle and Open Case Reserves

A distribution-free chain-ladder analysis of the real cumulative paid triangle (six accident years $\times$ seven development years; first age-to-age factor 1.116) gives an estimated ultimate of TRY 909.8 million against TRY 909.1 million already paid, implying an unpaid-claims reserve of TRY 689,034, about 0.08% of ultimate, so the triangle is almost fully run off. The Mack standard error [17] of the total reserve, including all cross-accident-year covariance terms [18], is TRY 541,623: only 0.06% of ultimate, but 78.6% of the reserve itself, which is what makes the reserving decision problem of Section 2.10 non-trivial despite the portfolio's maturity. The paid triangle excludes case reserves: at the valuation date (December 2020), 201 claim files remain open with case reserves of TRY 9.4 million (December-2018 prices), about 14 times the paid chain-ladder reserve, and 105 of these files (TRY 5.4 million) belong to accident years 2014–2017. Section 2.10 adds an incurred triangle and the historical run-off of case reserves, and Section 3.8 reconciles the

paid, incurred, and case-reserve bases; the analysis does not assess the adequacy of the insurer's booked reserves.

2.2. Computational Pipeline and Reproducibility

All calculations run in a seeded Python workflow of 26 modules with one configuration file (Python 3.12–3.13; NumPy 2.3–2.4, pandas 2.3–3.0, SciPy 1.16–1.17 [19], statsmodels 0.14, LightGBM 4.6.0, XGBoost 3.2–3.3, matplotlib 3.10; the pricing, premium, decomposition, and verification modules of this revision ran under Python 3.13.9 with NumPy 2.3.5, pandas 2.3.3, SciPy 1.16.3, statsmodels 0.14.5, and XGBoost 3.3.0); one command regenerates every table and figure from the administrative files. Each analysis module writes a machine-readable manifest, and a synthesis stage checks eleven consistency conditions—among them the temporal isolation of test data, the absence of leakage in the capital simulator, recovery in the positive control, reuse of the selected tail threshold, and the adequacy of the Monte Carlo tail sample—before consolidating the results. In addition, the headline results were re-derived along code paths independent of the pipeline's own implementation: a matrix-form chain-ladder and Mack implementation, the GLMs refitted with statsmodels, a class-based Kendall estimator, a simple policy-cluster bootstrap, a Lorenz-curve Gini, and a separate count-coupled simulator with empirical margins; all 32 checks agree within their stated tolerances (Section S6). The complete analysis code is openly available on GitHub and archived at Zenodo (https://doi.org/10.5281/zenodo.23894833; see the Data Availability Statement); aggregate derived data, including the verification table, are in Data S1. Generative artificial intelligence tools (Claude Fable 5.1, Anthropic; ChatGPT GPT-5, OpenAI) were used under the authors' direction for drafting and language editing, assistance in developing and checking the analysis and simulation code, and discussion and verification support for the formal results, but not to generate or alter data. The authors ran all analyses and simulations and verified all output.

2.3. Construction-Induced Frequency–Severity Association

Two constructions can make a frequency–severity association appear where no dependence exists: correlating the claim count with the total claim amount when only policy aggregates are available (the count–sum property), or with the average claim amount (aggregation-induced rank association). The first is elementary; the second is the construction behind copula models of summarized claims [11,12,15] and is the object of this section. A third practice, reusing the count's uniform, or its rank, to select severity quantiles in a compound simulation, is an implementation error rather than a measurement problem; its consequences (comonotonicity with the count, an upward shift in the realized severities, and an increase in the simulated loss in the increasing convex order) are stated and proved in Text S1, Section S1, and enter the paper only as step 5 of the protocol below.

2.3.1. Count–Sum Property

Let N be a policy's claim count with $0 < \mathrm{Var}(N) < \infty$, and let the severities $Y_1, Y_2, \ldots$ be independent and identically distributed (i.i.d.), independent of N, with $\mu_Y = \mathrm{E}(Y_1) > 0$; set $S = \sum_{k=1}^{N} Y_k$. Then $\mathrm{E}(S|N) = N\mu_Y$ and $\mathrm{Cov}(N,S) = \mu_Y \mathrm{Var}(N) > 0$, and for nonnegative severities the population Kendall coefficients $\tau_a(N,S)$ and $\tau_b(N,S)$ are positive, also conditionally on $\{N \geq 1\}$ (Text S1, Proposition S1 and Corollary S1). A positive count–sum association is thus an arithmetic property of the construction, not evidence about the claims process; the conditional version applies to the naive measure of Sections 3.1 and 3.2, which uses policies with $N \geq 1$.

2.3.2. Aggregation-Induced Rank Association

Remark 1 (Mechanism behind the count–average association)**.** *Under the model above,* $\mathrm{E}(S/N|N) = \mu_Y$ *on* $\{N \geq 1\}$*, so no linear association arises. With* $\bar{Y}_n$ *the mean of n i.i.d. severities and* $\pi_n = \Pr(N = n|N \geq 1)$*, the argument of Corollary S1 gives* $\tau_a(N, S/N \mid N \geq 1) = 2\sum_{1\leq n_1<n_2} \pi_{n_1}\pi_{n_2}\,\mathrm{E}[\mathrm{sgn}(\bar{Y}_{n_2} - \bar{Y}_{n_1})]$ *for independent* $\bar{Y}_{n_1}$ *and* $\bar{Y}_{n_2}$*. Each term vanishes for symmetric* $\bar{Y}_n$*, so the count-dependent variance alone creates no artifact. For right-skewed severities, however, the median of* $\bar{Y}_n$ *typically lies below* μ_Y *and rises toward it as its skewness* $\gamma_Y/\sqrt{n}$ *decays (*γ_Y *being the skewness of* Y_1*), so an average of more claims tends to exceed an average of fewer. For the lognormal fitted to the observed severities, Monte Carlo evaluation gives positive terms for all* $1 \leq n_1 < n_2 \leq 6$ (0.17 *for* $(n_1,n_2) = (1,2)$)*, whereas pairs involving seven or more claims carry a combined weight below* 10^{-3}*, far less than the weighted* (1,2) *term* ($2\pi_1\pi_2 \times 0.17 \approx 0.04$)*; hence the rank association is positive under independence. The* $1/N$ *variance scaling emphasized by Oh et al.* [15] *is therefore necessary but not sufficient for a rank artifact: the artifact is driven by skewness, and its size depends on the severity distribution and the claim-count composition.* Section 3.2 *evaluates the expression with the portfolio's own* π_n *and severities and compares it with a permutation null in which frequency and severity are exactly independent.*

2.3.3. Controlled Experiment and Diagnostic Protocol

To quantify both constructions under a known truth, we simulate casco-like portfolios from a common-latent-risk generator: a policy-level $Z \sim \mathcal{N}(0,1)$ drives the Poisson count ($\lambda = 0.35$) through $U = \Phi(Z)$, with Φ the standard normal distribution function, and each lognormal severity through a Gaussian-copula link with correlation

$$\rho = \sin(\pi\tau/2), \tag{1}$$

so the latent risk and each of its claim severities have Kendall's τ exactly, whereas claim-level measurement attenuates the observable count–severity association (Section 2.5). We sweep $\tau \in \{0, 0.05, 0.10, 0.20\}$ and portfolio sizes $n \in \{1000; 5000; 25{,}000; 100{,}000\}$, with 100, 33, 20, and 20 replications per (τ,n) cell, respectively. The consequence of reusing the count uniform in the severity generator is quantified in the same framework (30 replications of 20,000 policies without injected dependence; Text S1, Section S1).

2.4. Dependence Axes and Inference

2.4.1. Estimands and Dependence Axes

The primary frequency–severity evidence is a covariate-adjusted regression of log claim severity on the policy's claim count [8,9], with accident-year fixed effects, log earned exposure (the count depends on exposure), and the rating factors of Section 2.1 as potential confounders of both frequency and severity (step 3 of Algorithm 1). The claim-level specification is

$$\log Y_{ik} = \beta_0 + \beta_N N_i + \sum_{a=2015}^{2019} \gamma_a\, \mathbf{1}\{A_{ik} = a\} + \delta_0 \log e_i + \mathbf{x}_i^{\top}\boldsymbol{\delta} + \varepsilon_{ik}, \tag{2}$$

where Y_{ik} is the real severity of claim k on policy i in the severity sample, A_{ik} its accident year (2014 as the reference category), N_i the policy's claim count, which counts all linked claim files, and e_i the policy's earned exposure. The vector $\mathbf{x}_i$ collects the rating covariates: vehicle age, insured age, horsepower, engine displacement, and the nominal insured vehicle value, each standardized and entering linearly, together with dummy-coded gender (female and corporate; reference male) and body type (non-modal; reference modal). The rating factors are included as potential confounders: each plausibly affects both claim frequency and claim size, so omitting them would let risk-profile composition masquerade as a count

effect. The specification is deliberately conservative, since the target is the count coefficient rather than maximal predictive power; the pricing models of Section 2.8 relax the functional form where prediction is the objective. The count coefficient β_N (full results in Appendix A, Table A1) measures the direct frequency effect on severity, net of the observed risk profile and free of the count–sum and count–average constructions. Unlike the log-link severity models of Section 2.8, which predict $\log E[Y]$ from factors known at underwriting and so exclude the count, the regression conditions on the count and targets $E[\log Y]$, which is less sensitive to the heavy tail. Were the count added to a log-link model, the two count coefficients would coincide if the distribution of the log-scale error ε_{ik} did not depend on the count, because the two scales then differ only by the constant $\log E[e^{\varepsilon_{ik}}]$, which is absorbed in the intercept; they diverge when the dispersion varies with the count.

The rank-based counterpart of β_N, the frequency–severity (FS) axis, pairs the policy's claim count with each of its claim severities at claim level and is the capital simulator's calibration target (Section 2.7); the settlement-lag axis pairs each claim's mean payment delay (payment year minus accident year, averaged over the claim's payment records) with its severity and concerns reserve risk.

2.4.2. Inference: Cluster Bootstrap, Power, and Equivalence

All Kendall coefficients are tie-corrected τ_b [20], computed with scipy.stats.kendalltau [19] on the raw, tied data; each clean axis involves a heavily tied variable, the claim count on the FS axis and the settlement delay on the lag axis. Infinitesimal jitter is applied only where continuous margins are formally required—the probability-integral transforms (PITs) underlying copula estimation and two secondary tests—and never to a reported Kendall coefficient. With a discrete, heavily tied margin, rank-based copula inference depends on how the ties are broken [7]: on the FS axis, the maximum pseudo-likelihood estimate (MPLE) of the Gaussian copula parameter ranges from $\hat{\rho} = -0.0181$ to -0.0122 across 50 jitter seeds, and the bootstrap below redraws the jitter in every replicate so that its interval includes this variability. Because claims are clustered within policies, all primary confidence intervals (CIs) use a policy-cluster bootstrap [21] (policies resampled with all their claims; $B = 500$), with the i.i.d. bias-corrected accelerated bootstrap [22] as a check. A cross-fitted Vuong test [23] and Cramér–von Mises independence diagnostics [24] with Holm-adjusted p-values [25] are secondary evidence only, because copula inference for continuous margins is unreliable with heavily tied data [7] (Section S2).

Power is reported as the minimum detectable effect (MDE) of a two-sided 5% test with 80% power, $(z_{0.975} + z_{0.80})$ se ≈ 2.80 se, at two levels. The analytic value assumes independent observations and serves as an i.i.d. benchmark: with $n = 214{,}283$ claims and the observed ties, the exact null standard error of τ_b [20] is 0.00170 on the FS axis and 0.00169 on the settlement-lag axis (0.00144 for untied margins), giving MDEs of 0.0048 and 0.0047. The cluster-adjusted MDE, derived from the policy-cluster bootstrap standard error, is 0.0045 on the FS axis and 0.0049 on the settlement-lag axis; the corresponding design effects (variance ratios) relative to the tie-corrected benchmark are 0.89 and 1.09, whereas the larger ratios against the untied formula (1.24 and 1.50) mainly reflect ties rather than within-policy clustering.

Because the substantive claim is one of equivalence rather than a failure to reject, we apply the two one-sided tests (TOST) procedure [26] to $\hat{\tau}_{FS}$. For a margin $\Delta > 0$ the statistics are $z_- = (\hat{\tau}_{FS} + \Delta)/\widehat{se}$ and $z_+ = (\Delta - \hat{\tau}_{FS})/\widehat{se}$, with $\widehat{se}$ the policy-cluster standard error, and equivalence within Δ is declared when both exceed the 95% quantile of the standard normal distribution, that is, when the 90% interval $\hat{\tau}_{FS} \pm 1.645\ \widehat{se}$ lies inside $(-\Delta, \Delta)$; the smallest such margin is $|\hat{\tau}_{FS}| + 1.645\ \widehat{se}$. No external threshold for an economically relevant Kendall's τ exists; for this analysis, added after the primary one, we set the

reference margin $\Delta = 0.02$, more than four times the cluster-adjusted MDE and conservative on the capital scale (Section 3.5), and check the verdict over margins from 0.005 to 0.20 (Section S2).

2.5. Positive Control

A pipeline that finds no dependence should be shown capable of finding it. The positive control tests whether the measurement chain of Section 2.4 detects injected dependence and how strongly it understates it, validating the instrument, not the portfolio: it is synthetic by design and establishes the internal validity of the measurement chain, its ability to detect and quantify dependence injected into portfolios with the fitted marginals, not external validity across lines of business. With the latent-risk generator of Section 2.3 and the fitted marginals (negative binomial frequency, lognormal severity; Section 2.6), we generate 30 synthetic portfolios of 200,000 policies for each injected Kendall's $\tau \in \{0,0.10,0.25,0.40\}$ between the policy's latent risk and its claim sizes, and rerun the identical pipeline. We report mean recovery and the empirical Type-I error and power, with Clopper–Pearson intervals, of an analytic per-claim test of $\tau = 0$ (i.i.d. null standard error without tie correction, not the policy-cluster inference).

The capital criterion compares the policy-level, not portfolio-level, 99.5% loss quantile with and without the injected dependence, because the policy loss is the object through which a frequency–severity link first manifests; the portfolio-aggregate counterpart is the capital experiment of Section 2.7, and the two objects are kept distinct. Because the latent risk also drives the count, claims arise only on high-risk policies, shifting the claim-severity distribution upward. We therefore use two independent arms: one drawn from the unconditional severity margin (the total effect, including this shift) and one that permutes the dependent arm's realized claim severities across all claims of the portfolio, holding the claim-level severity distribution fixed exactly (the dependence effect, the counterfactual of a simulator calibrated to observed claims); the permutation breaks both the link between claim sizes and the count and the co-movement of claim sizes within a policy, both of which belong to the injected dependence. Three criteria are applied: monotone recovery; a capital effect above 10% at $\tau = 0.40$; and one below 2% in absolute value at $\tau = 0$. The thresholds were prespecified for the total effect; the margin-preserving comparison, to which they are also applied, was added in revision. Recovery is attenuated because claim-level measurement conditions on $N \geq 1$, range-restricting the latent risk among claims (Section 3.3).

2.6. Marginal and Copula Models

We model severity in two parts: a parametric body and an extreme-value tail. Among lognormal, gamma, and Weibull candidates fitted by maximum likelihood, the body is selected by the Akaike information criterion (AIC). The tail follows the peaks-over-threshold generalized Pareto distribution (GPD) [27–29]. Rather than fixing the customary 90th-percentile threshold, the threshold is selected on the candidate quantile grid {0.85, 0.90, 0.925, 0.95, 0.975} by minimizing the Kolmogorov–Smirnov (KS) distance between exceedances and the fitted GPD, subject to a minimum of 500 exceedances, a data-driven implementation of the threshold-stability principle; the selected threshold quantile is stored and reused by the capital simulation, so that the body–tail splice is internally consistent. Claim frequency is fitted by Poisson and negative binomial likelihoods, with the negative binomial dispersion estimated by maximum likelihood, and selected by AIC.

On each dependence axis, PITs are computed by ranks (with tie jitter) and one-parameter copulas—Joe, Frank, Gumbel, Clayton, survival Clayton, and Gaussian—are estimated by MPLE, with inversion of Kendall's τ (sign-aware for negative dependence)

as the starting value. The independence copula is included as an explicit candidate, and families are compared by AIC and the Bayesian information criterion (BIC); upper- and lower-tail dependence coefficients are reported with the fits. The selected family's policy-cluster bootstrap CI redraws the jitter, recomputes the PITs, and re-estimates the MPLE in every resample (full-size resamples of policies, each bringing all its claims; $B = 500$; percentile interval), so that point estimate and interval share one estimator.

2.7. Capital Experiments: Paired and Leakage-Free

2.7.1. Paired Simulation Design

The aggregate loss $S = \sum_i S_i$, $S_i = \sum_{k=1}^{N_i} Y_{ik}$, of a portfolio of 100,000 policies is simulated from the fitted negative binomial frequency and the spliced body–GPD severity in 20,000 paired replications (four independent seed blocks). In the independent arm, policy i has count $N_i = F_N^{-1}(U_i)$ and claim sizes $Y_{ik} = F_Y^{-1}(p_{ik})$, with independent uniforms U_i and p_{ik} and F_Y the fitted splice. The dependent arm keeps N_i and p_{ik} and couples the claim sizes to the count's own uniform through an explicit Gaussian copula,

$$z_{ik} = \rho\,\Phi^{-1}(U_i) + \sqrt{1-\rho^2}\,\Phi^{-1}(p_{ik}), \qquad Y_{ik}^{\mathrm{dep}} = F_Y^{-1}\{H_\rho(\Phi(z_{ik}))\}, \tag{3}$$

where H_ρ is the claim-weighted distribution function of $V_{ik} = \Phi(z_{ik})$,

$$H_\rho(v) = \frac{\mathrm{E}\{N_i \Pr(V_{ik} \le v \mid U_i)\}}{\mathrm{E}(N_i)} = \frac{1}{\mathrm{E}(N_i)} \sum_{n \ge 1} n \int_{F_N(n-1)}^{F_N(n)} \Phi\left\{ \frac{\Phi^{-1}(v) - \rho\,\Phi^{-1}(w)}{\sqrt{1-\rho^2}} \right\} dw,$$

which is not uniform when $\rho \neq 0$, because claims are count-weighted and the count is coupled to V_{ik} through U_i (Gauss–Legendre quadrature; Section S3). This design excludes both distortions of shared-uniform leakage (Text S1, Section S1). There is no leakage: U_i is never reused as a severity uniform (Proposition S2) and affects claim sizes only through a copula of explicit strength, so the arms coincide at $\rho = 0$. Nor is the severity margin shifted: the monotone correction H_ρ makes the count-weighted V_{ik} uniform at claim level, so both arms have exactly the fitted claim-level severity distribution, while claim-level ranks are unchanged. The coupling thus alters only how claim sizes depend on the count and, through the shared U_i, on one another within a policy; pairing isolates it from Monte Carlo noise.

Each arm is summarized by $\mathrm{VaR}_q(S)$ and $\mathrm{TVaR}_q(S)$ at $q = 0.95$, 0.99, and 0.995 and by the capital distance $K_q = \mathrm{VaR}_q(S) - \mathrm{E}(S)$; we write $K = K_{0.995}$. Changes are reported as $X^{\mathrm{ind}}/X^{\mathrm{dep}} - 1$ (positive: independence gives the higher figure); as the expected loss dominates the aggregate VaR, changes are far larger relative to K. The 99.5% quantities are Solvency-II-style capital proxies [30], not the solvency capital requirement (SCR), which also covers other risk modules, diversification, and a one-year view. Monte Carlo uncertainty is quantified by paired bootstrap 95% CIs ($B = 1000$), with per-block seed-sensitivity checks and a split-sample Kupiec test [31] of quantile estimation (Section S3).

2.7.2. Calibration, Sensitivity Sweep, and Materiality

The coupling is calibrated so that the simulated claim-level $\tau_b(N,Y)$ equals the observed $\hat{\tau}_{\mathrm{FS}} = -0.0065$ (the unadjusted association, not the covariate-adjusted β_N). By (1), ρ fixes the latent Kendall's $\tau_{\mathrm{lat}} = (2/\pi)\arcsin\rho$ between U_i and V_{ik}, which claim-level measurement attenuates (Section 2.5); simulated portfolios of five million policies at $\tau_{\mathrm{lat}} \in \{-0.05, -0.10, -0.20, -0.40\}$, fitted as $\tau_b = c\,\tau_{\mathrm{lat}} + c_3\,\tau_{\mathrm{lat}}^3$, give the attenuation slope $c = \tau_b/\tau_{\mathrm{lat}} = 0.644$ at the origin, hence

$$\tau_{\mathrm{lat}}^* = \hat{\tau}_{\mathrm{FS}}/c = -0.0102, \qquad \rho^* = \sin(\pi\tau_{\mathrm{lat}}^*/2) = -0.0160. \tag{4}$$

At ρ^* the exact population value of the claim-level τ_b is -0.0065, the calibration target; a simulation check on ten million policies gives -0.0057 (Monte Carlo standard error 0.0007). The simulator thus absorbs the attenuation by construction and uses neither a direct inversion of $\hat{\tau}_{FS}$ nor the MPLE of Section 2.6. A copula-free sensitivity draws each claim's severity rank from the empirical distribution of severity PITs within the policy's count class (1, 2, 3, 4+), with the analogous margin correction. A sweep over $\tau_{lat} \in \{\pm 0.05, \pm 0.10, \pm 0.20, \pm 0.40\}$ (20,000 paired replications each, same random streams) locates where the capital effect would become material; a first-order analytic check predicts the change in K from the change in the standard deviation of S implied by count-class severity moments (Section S3). Absent a regulatory threshold, we take a 5% change in K as the materiality reference, a decision criterion of this study rather than a regulatory or universal threshold, and report, from the analytic curve, the positive claim-level association at which assuming independence would understate K by 1% and by 5%.

2.8. Pricing Models and Temporal Validation

2.8.1. Temporal Design and Response

Pricing models are evaluated only on the test years 2017–2018 (295,853 policies). All models, generalized linear models (GLMs) and boosted models alike, are fitted on the same 356,846 policies of 2014–2015, and 2016 (166,090 policies) serves only for validation: early stopping, the Tweedie variance power, model selection, the level adjustment, and recalibration. Numeric covariates are standardized with the 2014–2015 means and standard deviations, applied unchanged to later years, so no information from the validation or test years enters any fitted model or transformation; a sensitivity refits the GLMs on 2014–2016 (Section S4). The response is the per-policy ultimate claim cost in real terms: paid amounts plus the case reserve still open at the valuation date (December 2020), floored at zero; case reserves recorded earlier for claims closed by that date are not added, because their outcome is already reflected in the paid amounts, and later development is negligible. Severity models use this amount per claim; frequency-only predictions are scaled by a constant mean severity, which leaves their Gini index unchanged.

2.8.2. Model Set, Selection, and Evaluation

A 2 × 2 design crosses model structure (direct Tweedie pure premium [32,33] versus two-stage Poisson frequency × log-link gamma severity [34]) with learner type (GLM versus gradient-boosted machine, GBM [35]). The two-stage GLM combines the Poisson frequency GLM with a log-link gamma severity GLM fitted on claimant policies with claim counts as weights, the standard multiplicative tariff structure of non-life GLM pricing; the two-stage GBM combines a Poisson frequency GBM with a gamma severity GBM. The Tweedie GLM power is selected in 2016 by profile likelihood with the exact density [36] over $p \in \{1.10, 1.15, \ldots, 1.90\}$; $\hat{p} = 1.60$ is the only grid value inside an approximate 95% interval (Section S4). Poisson and negative binomial GLMs with a log-exposure offset serve as frequency-only baselines, and LightGBM [37] and XGBoost [38] (Tweedie objectives with $p = 1.5$; single-threaded, fixed seeds) as package baselines. Apart from the number of trees and p, all settings were fixed a priori; Appendix B (Table A2) gives the full specification of every model.

Discrimination is measured by the normalized pure-premium Gini index, with paired-bootstrap percentile intervals ($B = 500$) for all model differences and 2 × 2 contrasts. The model carried forward to the calibration, premium, and capital calculations is the one with the highest validation-year (2016) Gini; test-year results are reported for every model and are never used for selection. Calibration is assessed by the predicted-to-observed ratio and

by decile- and policy-level calibration slopes (ideal value one; Section S4), before and after isotonic recalibration learned on 2016 [39].

Because a model fitted on 2014–2015 carries the claim-frequency level of those years, its predictions are brought on level before they enter the premium target, as pricing practice does each year. Three scalar factors are computed without any test-year information and compared in Section 3.6: the validation-level factor, the ratio of realized to predicted pure premium on 2016; the frequency-trend factor, the log-linear trend of the claim frequency per policy-year over 2014–2016 extrapolated from the exposure-weighted center of the fit years to that of the test years; and their standard combination, which first re-levels the predictions to 2016 and then trends them from 2016 to the center of the test years. A scalar factor changes the level of the predictions only, so the Gini index, the calibration slopes, and all relative diagnostics are unaffected; the combined factor is the one carried forward.

2.9. RORAC-Target Technical Premium and Premium Principles

The capital loading of the selected model's expected loss $m(x)$ assumes that each test policy's loss S_i is compound Poisson with intensity $\lambda_i = m(x_i)/\mathrm{E}[Y]$ and portfolio claim size Y, so its process variance is $V_i = m(x_i)\,\mathrm{E}[Y^2]/\mathrm{E}[Y]$ (empirical claim-file moments). The portfolio capital proxy SCR is the 99.5% quantile minus the mean of the independent-sum aggregate $S = \sum_i S_i$, obtained by a Cornish–Fisher expansion [40] of the compound cumulants and checked by Monte Carlo simulation. A negative binomial sensitivity recomputes it with $\mathrm{Var}(S_i) = \mathrm{E}[N_i]\mathrm{Var}(Y) + D\,\mathrm{E}[N_i]\,\mathrm{E}[Y]^2$, where $D = 1.233$ is the variance-to-mean ratio of the claim count or its policy-bootstrap 97.5th percentile (Section S4). The proxy is allocated as $A_i = \mathrm{SCR}\,V_i/\sum_j V_j$, its exact Euler allocation [41] (Remark 2).

Remark 2 (Allocation exactness)**.** *Let policy losses be independent, each S_i compound Poisson with intensity λ_i and a common claim-size distribution with finite moments of the orders used. Then all cumulants are proportional to the expected claim frequency: $\kappa_j(S_i) = \lambda_i\,\mathrm{E}[Y^j]$ and $\kappa_j(S) = \Lambda\,\mathrm{E}[Y^j]$, with $\Lambda = \sum_i \lambda_i$. Let $\varrho(S)$ be the Cornish–Fisher 99.5% quantile of S, truncated at any order, minus* $\mathrm{E}[S]$*, viewed as a differentiable function of the cumulants (κ_2,κ_3,...). Scaling the loss, $S \mapsto cS$ with $c > 0$, maps $\kappa_j \mapsto c^j\kappa_j$ and $\varrho \mapsto c\varrho$ (positive homogeneity of degree one in the loss, not in the cumulants), so differentiating at $c = 1$ gives the Euler identity $\sum_j j\,\kappa_j\,\partial\varrho/\partial\kappa_j = \varrho(S)$. The Euler contribution of policy i scales its own summand within S: $S + \varepsilon S_i = (1+\varepsilon)S_i + \sum_{l\neq i} S_l$ has cumulants $\kappa_j(S) + \{(1+\varepsilon)^j - 1\}\,\kappa_j(S_i)$ by independence, hence $A_i = \partial_\varepsilon\,\varrho(S+\varepsilon S_i)|_{\varepsilon=0} = \sum_j j\,\kappa_j(S_i)\,\partial\varrho/\partial\kappa_j = \lambda_i\sum_j j\,\mathrm{E}[Y^j]\,\partial\varrho/\partial\kappa_j$, proportional to λ_i, and $\sum_i A_i = \sum_j j\,\kappa_j(S)\,\partial\varrho/\partial\kappa_j = \varrho(S)$ by the Euler identity. Thus $A_i = \varrho(S)\,\lambda_i/\Lambda$, the variance shares above, at every order; the approximation error lies in the Cornish–Fisher quantile, not in the allocation rule. The leading skewness correction, $(\gamma/6)(z_{0.995}^2 - 1)\,\sigma_S$, where σ_S and $\gamma = 0.038$ are the standard deviation and skewness of S, is about 1.4% of the capital proxy $\varrho(S)$, and the Cornish–Fisher capital lies 10% above the Monte Carlo check (*Section 3.7*).*

The RORAC-target premium and the resulting portfolio RORAC are

$$\pi^*(x) = \frac{m(x) + h\,A(x)}{1-e}, \qquad \mathrm{RORAC} = \frac{(1-e)\sum_i \pi^*(x_i) - \sum_i m(x_i)}{\mathrm{SCR}}, \tag{5}$$

with hurdle rate $h = 12\%$, expense ratio $e = 25\%$, and $A(x_i) = A_i$. Because $\sum_i A_i = \mathrm{SCR}$, the portfolio attains RORAC = h by construction; π^* is a target, the smallest aggregate premium attaining h under Euler-proportional loadings, not an optimum. The expected loss $m(x)$ is the on-leveled prediction of Section 2.8. Classical premium principles [42,43] and the cost-of-capital principle [30] distribute the same margin $h \cdot \mathrm{SCR}$ differently across policies; because every cumulant of a policy's loss is proportional to its expected loss under this model, four of the five principles induce identical risk-proportional loadings, and

the comparison is relegated to Text S1 (Section S4), together with the relative adequacy diagnostic of the collected premiums, whose recorded scope cannot be verified.

2.10. Reserve Bases and Confidence-Level Choice

The paid chain-ladder reserve and the case reserves open at the valuation date differ by an order of magnitude (Section 2.1.4); before any confidence level is chosen, the bases must be reconciled. Three bases are compared. (i) The paid basis: the chain-ladder reserve of the real cumulative paid triangle with its Mack standard error [17,18]. (ii) The incurred basis: the same method applied to the incurred triangle, whose cell for accident year i and development year j is the cumulative real paid amount through the calendar year $i + j$ plus the real case reserves open at the December snapshot of that year (closed claims leave the book); the outstanding liability is the incurred ultimate minus the amount paid to date. (iii) The run-off-adjusted basis: for each December snapshot from 2014 to 2019, the case reserves held on the then-open files are compared with what was subsequently paid on the same files up to the valuation date; the ratios from the mature snapshots (2014–2017, at least three years of run-off), pooled and by age of the file at the snapshot (under one year, one year, two years or more), are applied to the case reserves of the files open at the valuation date, matched by age.

On the paid and incurred bases, the outstanding amount X is given a lognormal predictive distribution F, moment-matched to the chain-ladder best estimate (BE) and the Mack standard error; since Mack's model does not imply lognormality, F is stress-tested against moment-matched gamma, Weibull, and inverse Gaussian alternatives and an over-dispersed Poisson bootstrap [44,45] (Section S5). Confidence-level reserves VaR_q and TVaR_q for $q \in \{75\%, 95\%, 99.5\%\}$ span an illustrative 75th-percentile benchmark (International Financial Reporting Standard 17 [46] sets no fixed level) and the Solvency-II-style 99.5% level [30], and the technical provision is BE plus a cost-of-capital (CoC) risk margin RM,

$$\mathrm{RM} = \mathrm{CoC} \sum_{t=0}^{T-1} \frac{\mathrm{SCR}_t}{(1 + r_\mathrm{f})^{t+1}}, \qquad \mathrm{SCR}_t = \mathrm{SCR}_0 \, \frac{O_t}{O_0}, \tag{6}$$

with standard Solvency II timing: O_t is the expected amount outstanding at the end of the run-off year t ($t = 0$ is the valuation date; $T = 7$), computed from the expected chain-ladder payments of the paid triangle by future calendar year with negative expected increments set to zero; $\mathrm{SCR}_0 = \mathrm{VaR}_{0.995}(X) - \mathrm{BE}$, an ultimate-view proxy that tends to overstate the one-year reserve-risk capital; CoC = 6%; and $r_\mathrm{f} = 0$ in constant prices. The cost-based choice of the confidence level through the newsvendor critical fractile, with its distributional sensitivity, is given in Text S1, Section S5.

3. Results

Sections 3.1–3.5 establish what survives clean measurement and what it implies for capital; Sections 3.6–3.8 report the secondary implications for pricing, the premium target, and the reserve bases.

3.1. Size of the Artifacts Under a Known Truth

Under exact independence (true $\tau = 0$; Figure 1), the count–sum construction returns a mean Kendall τ of +0.29 and the count–average construction +0.07, the median effect of Remark 1, whereas the clean per-claim measure stays centered at zero at every portfolio size (means between -0.0001 and +0.0020). At positive true τ the clean measure preserves the ordering of the scenarios but attenuates the injected value (0.032 at $\tau = 0.05$ and 0.067 at $\tau = 0.10$ for $n = 100{,}000$), as explained in Section 3.3. An analyst using the count–sum construction would thus report strong dependence in a portfolio without any, and neither

construction is evidence about the dependence actually present. Reusing the count uniform in the severity generator, the implementation error of Section 2.3 raises the simulated 99.5% VaR of the policy loss by 276% at zero true dependence (Figure 1c; Text S1, Section S1).

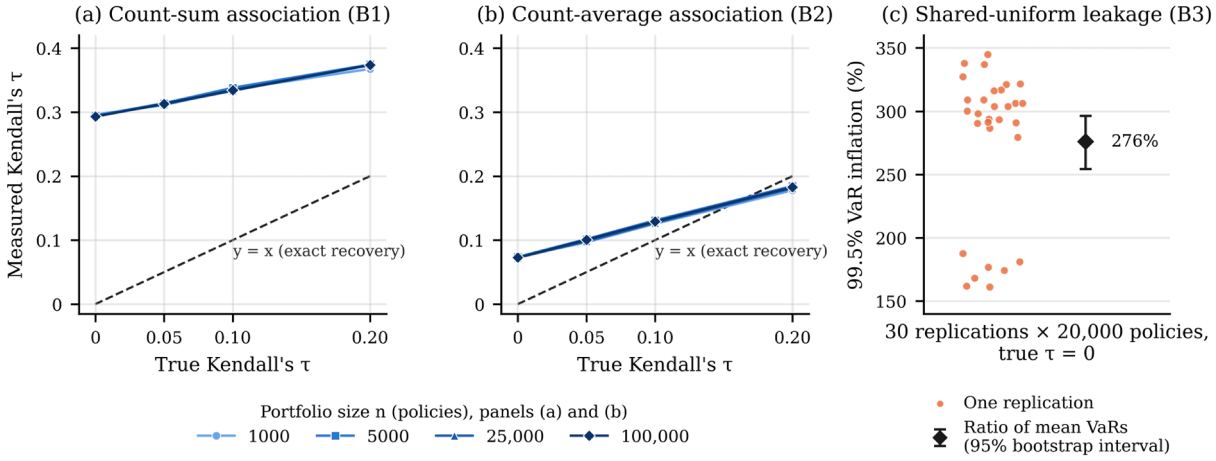

Figure 1. Artifacts under a known truth. (**a**,**b**) Mean Kendall's τ of count–sum and count–average pairs versus true τ, by portfolio size. (**c**) Simulated 99.5% VaR inflation by shared-uniform leakage at true $\tau = 0$: replications and ratio of mean VaRs with 95% bootstrap interval. In the panel labels, B1, B2, and B3 denote the count–sum, count–average, and shared-uniform constructions. VaR, value-at-risk.

3.2. Dependence Diagnostics on the Casco Portfolio

The portfolio reproduces the pattern of the controlled experiment (Table 2; Figure 2). Both naive measures are strongly positive: $\tau = +0.254$ for count against aggregate loss and +0.081 against mean severity. Read literally, the count–sum value would signal dependence as strong as the middle positive-control injection ($\tau = 0.25$; Section 3.3), but it stems from the construction (Section 2.3), not from the claims process.

Table 2. Dependence estimates for 214,283 claims from 818,789 policies. 95% CIs are policy-cluster bootstrap percentile intervals ($B = 500$). Kendall coefficients are τ_b on raw, tied data; the naive rows A and B are policy-level and have no interval. CI, confidence interval; FS, frequency–severity.

Axis	Construction	Estimate	95% CI	Interpretation
A (naive)	count × aggregate loss	$\tau = +0.254$	—	count–sum property; +0.275 under independence (Table 3)
B (naive)	count × mean severity	$\tau = +0.081$	—	aggregation-induced; +0.106 under independence
FS-adj	log-severity regression	$\beta_N = -0.0436$	[−0.0515,−0.0355]	primary; −4.3% per claim, covariate-adjusted (Table A1)
FS (clean)	count × per-claim severity	$\tau = -0.0065$	[−0.0098,−0.0035]	small; pricing/capital axis
Lag (clean)	payment delay × severity	$\tau = +0.0185$	[+0.0152,+0.0221]	weak; reserving axis

How much of each naive measure the construction accounts for is answered by a permutation null with the portfolio's own margins (Table 3; Figure 3): severities are permuted across the 214,283 claims, which makes frequency and severity exactly independent while preserving every policy's claim count and the severity distribution, and the three measures are recomputed over 200 replicates. Under exact independence, the count–average coefficient is +0.106 (95% range [0.103,0.109]), larger than the observed +0.081, and the count–sum coefficient +0.275 against the observed +0.254; the clean claim-level coefficient is centered at zero (0.0000, range [−0.003,+0.003]). The whole of the observed count–average association, and more, is therefore reproduced by the construction: a copula fitted to this pair would absorb a positive association that does not belong to the loss process, and

the residual, observed minus null, is negative (-0.025) and of the same sign as the clean coefficient, so the naive pair would even return the wrong sign of what dependence there is. When the severities are replaced by a symmetric distribution with the same mean and variance, so that the $1/N$ variance scaling of the average is retained but its skewness removed, the count–average null falls to -0.000 ([-0.004,+0.003]) while the count–sum null remains positive (+0.130): the rank artifact of the count–average pair is driven by skewness, as Remark 1 states, not by the variance scaling. The analytic expression of Remark 1, evaluated with the portfolio's claim-count shares ($\pi_1 = 0.80$, $\pi_2 = 0.16$, $\pi_3 = 0.03$) and the fitted lognormal, gives +0.105 on the τ_b scale (+0.108 with the empirical severities), in agreement with the permutation value. These numbers are specific to this portfolio's severity skewness and claim-count composition; the decomposition itself is the diagnostic.

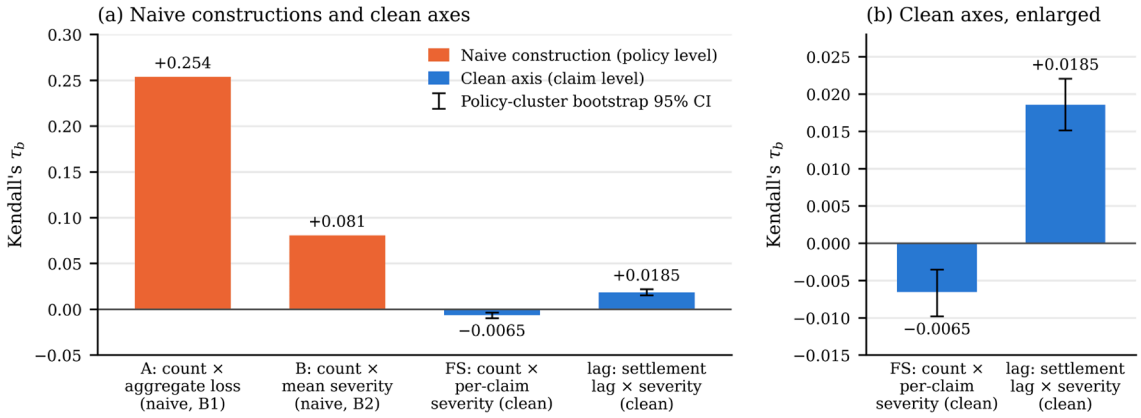

Figure 2. Naive and clean dependence measures in the casco portfolio (Kendall's τ_b). (**a**) All four measures on a common scale: A, count versus aggregate loss (count–sum); B, count versus mean severity (count–average); FS, count versus per-claim severity; lag, settlement lag versus claim severity. A and B are policy-level, FS and lag claim-level coefficients (B1 and B2 in the panel labels denote the count–sum and count–average constructions). (**b**) The two clean axes on an enlarged scale, with policy-cluster bootstrap 95% confidence intervals ($B = 500$). The regression coefficient β_N, not on the τ scale, is omitted.

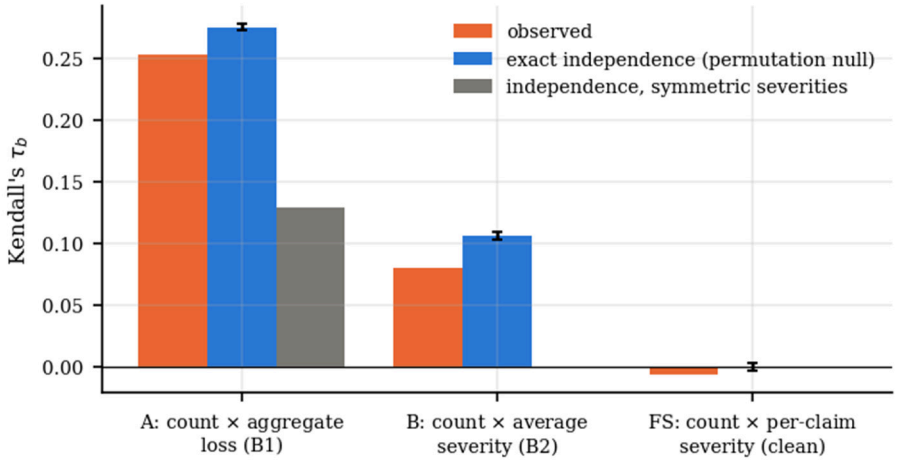

Figure 3. Observed naive and clean measures of the casco portfolio against their values under exact frequency–severity independence (permutation null, 2.5–97.5% range over 200 replicates) and under independence with symmetric severities (Table 3). The count–average artifact disappears when skewness is removed; the count–sum artifact does not. B1 and B2 in the axis labels denote the count–sum and count–average constructions.

Table 3. Decomposition of the naive measures of the casco portfolio into the part reproduced under exact frequency–severity independence and the residual (Kendall's τ_b; 214,283 claims on 171,822 claimant policies). Independence null: severities permuted across claims, claim counts and the severity margin preserved; mean and 2.5–97.5% range over 200 replicates. Symmetric-severity null: the same with severities drawn from a normal distribution with the observed mean and standard deviation (variance scaling of the average retained, skewness removed). Residual: observed minus independence-null mean.

Measure	Observed	Independence Null	Residual	Symmetric-Severity Null
A: count × aggregate loss (count–sum)	+0.254	+0.275 [0.273,0.278]	−0.022	+0.130 [0.125,0.134]
B: count × average severity (count–average)	+0.081	+0.106 [0.103,0.109]	−0.025	−0.000 [−0.004,+0.003]
FS: count × per-claim severity (clean)	−0.0065	+0.000 [−0.003,+0.003]	−0.0066	+0.000 [−0.004,+0.003]

After adjustment, the sign reverses, and the magnitude becomes small. In the covariate-adjusted regression (2) (Table A1), one additional claim on a policy is associated with 4.3% lower individual severity ($\beta_N = -0.0436$; policy-cluster 95% CI [−0.0515,−0.0355]). The claim-level rank estimate agrees: $\hat{\tau}_{\mathrm{FS}} = -0.0065$ [−0.0098,−0.0035] statistically detectable, small, and opposite in sign to the naive measures. Settlement lag is weakly associated with severity ($\hat{\tau}_{\mathrm{lag}} = +0.0185$ [+0.0152,+0.0221]), a reserving rather than a pricing matter; the association persists within accident years (claim-weighted mean of the accident-year coefficients +0.0270). Cross-fitted Vuong statistics favor independence on both axes and Cramér–von Mises diagnostics fail to reject it (p = 0.668 for FS and 0.140 for lag); on heavily tied margins, these secondary tests are far less sensitive than the tie-corrected rank coefficients (Section S2). Power is not limiting: $|\hat{\tau}_{\mathrm{FS}}| = 0.0065$ exceeds its cluster-adjusted MDE of 0.0045, the lag estimate is 3.8 times its MDE of 0.0049, and power at the observed values is 0.98 (FS) and above 0.99 (lag).

The equivalence test formalizes the verdict. With the reference margin $\Delta = 0.02$, both one-sided tests reject ($z_- = 8.39$, $z_+ = 16.55$; $p < 10^{-16}$): the 90% interval [−0.0092,−0.0039] lies inside $(-\Delta, \Delta)$, and the smallest margin at which equivalence holds is 0.0092; on the grid from 0.005 to 0.20, equivalence is declared for every margin of at least 0.01 (Section S2). The coefficient is thus both distinguishable from zero and significantly within the equivalence region. Three quantities are therefore kept apart. The conditional effect β_N is real and not negligible, but of limited aggregate impact: claims on multi-claim policies are 35.7% of all claims, and applying the factor $e^{\beta_N(k-1)}$ to each claim on a k-claim policy lowers expected losses by about 1.7% when claims are weighted by their observed severities (2.1% with equal weights); as the count is known only after the fact, it is a compositional feature rather than an a priori rating factor, although it enters experience rating such as bonus–malus schemes [2]. The rank association $\hat{\tau}_{\mathrm{FS}}$ is small relative to Δ, and its capital consequence is immaterial (Section 3.5).

3.3. Positive Control: Recovery of Known Dependence

Over 30 portfolios of 200,000 policies per injection level (Figure 4; ± values are between-portfolio standard deviations), recovery is monotone and stable. For injected τ of 0.10, 0.25, and 0.40, the mean claim-level latent estimates are 0.059 ± 0.003, 0.150 ± 0.003, and 0.252 ± 0.003, and the per-claim estimates (τ_b of count and severity) are 0.065, 0.164, and 0.275; without injection, the latent estimate is −0.0006 ± 0.0023. The Type-I error of the per-claim test of $\tau = 0$ is 0.033 (exact 95% interval [0.001,0.172]; nominal 0.05), and every positive scenario is detected in all 30 replications (Clopper–Pearson lower power bound 0.884).

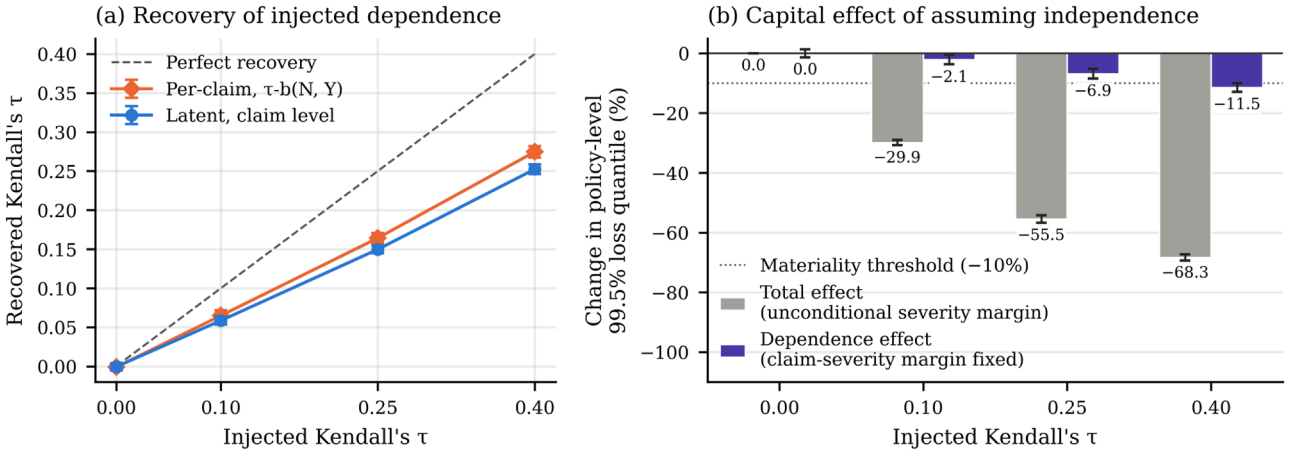

Figure 4. Positive control (30 portfolios of 200,000 policies per level). (**a**) Recovered latent and per-claim τ versus injected τ. (**b**) Policy-level 99.5% loss-quantile change from assuming independence: total and dependence effects. Error bars: ± 1.96 between-portfolio standard deviations.

The attenuation has an identifiable source. Claim-level measurement conditions on $N \geq 1$: severities exist only for policies whose latent risk exceeds the zero-claim threshold $\Phi^{-1}\{\Pr(N = 0)\} = 0.81$ (79% of policies have no claim), so the latent risk is range-restricted in the claim sample. To separate this range restriction from count ties, Kendall's τ between the policy latent and a severity latent is computed over all policies (a severity latent is drawn for every policy), over claimant policies only, and at claim level (one pair per claim); none of these measures involves the count. Over all policies, the latent coefficient equals the injected value (ratio 1.00 at every level); over claimant policies, it falls to 0.44–0.50 of the injected value, and at claim level it retains 0.59–0.63. Count ties are therefore not the source of the attenuation. The discreteness of the count adds only a ceiling: for a continuous severity and the observed claim-count composition,

$$\tau_b \leq \sqrt{1 - \sum_k q_k^2} = 0.72, \qquad \tau_a \leq 1 - \sum_k q_k^2 = 0.516,$$

where q_k is the share of claims on policies with count k; the bounds are attained only when count and severity are comonotone, an instance of the general fact that the margins restrict the attainable range of a dependence measure [13]. The ceiling is not binding: the tie-corrected per-claim coefficient is slightly larger than the latent one (0.275 against 0.252 at $\tau = 0.40$). The per-claim recovered-to-injected ratio is stable across designs: 0.65–0.69 in the positive control and 0.64–0.67 in the stylized generator of the artifact experiment at $n = 100{,}000$ (0.54–0.70 across all portfolio sizes, the wider spread at small n being mostly Monte Carlo noise). Both designs use lognormal severities, but rank coefficients are invariant to strictly increasing transformations of the severity, so, for a given latent-risk generator, the ratio depends on the marginals only through the count distribution, which differs between the two designs (Poisson with 70% and negative binomial with 79% claim-free policies). The capital experiment absorbs the attenuation by calibrating its latent coefficient directly to the observed claim-level coefficient (Section 2.7).

Against the independent arm drawn from the unconditional severity margin, assuming independence understates the policy-level 99.5% loss quantile by 29.9% $\pm$ 0.5, 55.5% $\pm$ 0.6, and 68.3% $\pm$ 0.5 at the three positive levels (0.0% at $\tau = 0$). Most of this total effect is the severity shift in Section 2.5: the dependent arm's mean claim severity is 1.40, 2.17, and 3.06 times its unconditional value. With the claim-severity distribution held fixed, the understatement (the dependence effect) is 2.1% $\pm$ 0.8, 6.9% $\pm$ 0.8, and 11.5% $\pm$ 0.7 (0.04% $\pm$ 0.7 at $\tau = 0$). At the strongest injection, ignoring dependence would thus set

this quantile about 11% too low even with an exact severity distribution, and at about one third of its true value if the shift were also missed. All criteria are met, with the thresholds prespecified for the total effect applied to the margin-preserving comparison added in revision, so the small portfolio estimate reflects the data, not the diagnostic: recovery is monotone, the dependence effect at $\tau = 0.40$ exceeds the 10% threshold (Monte Carlo 95% interval [11.2%,11.7%]; 29 of 30 portfolios), and the null effect stays below 2%.

3.4. Marginal Distributions and Copulas

AIC selects a lognormal severity body over gamma and Weibull (Figure 5a), and the threshold rule selects the 92.5th percentile (TRY 12,500; 16,072 exceedances; KS distance 0.0046). The GPD shape is $\hat{\xi} = 0.308$ (95% CI [0.289,0.330]), a tail index of about 3.24: heavy, but with finite second and third moments (exceedance-bootstrap 95% CI of the tail index [3.03,3.46]). Along the candidate grid, the estimated shape falls from 0.402 to 0.299 and is already stable around the selected threshold (Figure 6a); at the conventional 90th percentile it would have been 13% higher ($\hat{\xi} = 0.349$, a tail index of 2.87). Above the selected threshold, the empirical mean excess is close to linear, and the line implied by the fitted GPD stays within its 95% pointwise band (Figure 6b). The choice matters little economically: across all candidate thresholds, a compound-Poisson capital proxy (99.5% quantile minus mean, by a Cornish–Fisher approximation for one year's claim volume of about 42,900 claims) stays between 3.2% and 3.9% of expected losses. Claim counts are over-dispersed; the negative binomial is selected ($r = 1.129$). Figure 5 shows the selected lognormal body and the negative binomial count fit against its Poisson alternative.

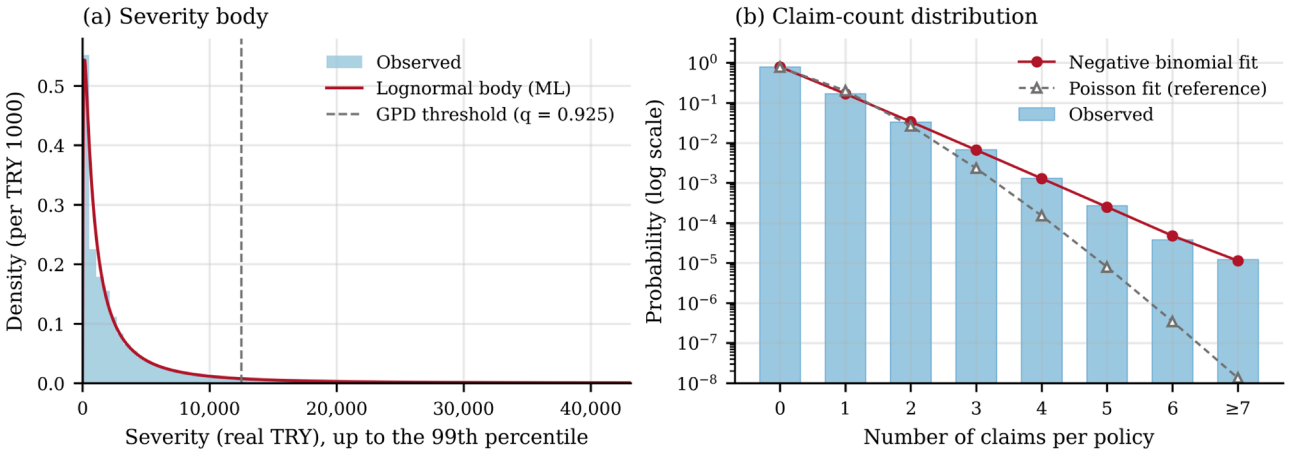

Figure 5. Marginal fits. (**a**) Claim severities in real (December 2018) TRY up to the 99th percentile (TRY 43,152), with the fitted lognormal body density; the dashed line marks the GPD threshold ($q = 0.925$, TRY 12,500), above which the GPD tail applies (Figure 6). (**b**) Claim counts per policy (last category pooled, ≥7; logarithmic scale) against the maximum-likelihood negative binomial fit ($r = 1.129$) and, for reference, the Poisson fit with the same mean. GPD, generalized Pareto distribution; TRY, Turkish lira.

The copula fits (Table 4; ledger in Table 5) leave little dependence to model: a Gaussian copula with $\hat{\rho} = -0.0177$ on the FS axis, without tail dependence, and a Joe copula with $\hat{\theta}_{\mathrm{J}} = 1.0461$ on the lag axis, with a parameter just above the independence value $\theta_{\mathrm{J}} = 1$ and weak upper-tail dependence ($\lambda_U = 0.060$). AIC and BIC prefer them to independence, and their cluster intervals exclude the independence value: with 214,283 claims, even $\hat{\tau}_{\mathrm{FS}} = -0.0065$ is resolved, since on the FS axis twice the log-likelihood gain, 67.4, far exceeds the BIC penalty of $\log n \approx 12.3$ for the single parameter. The Kendall's τ implied by a fitted copula refers to the jittered transforms and is not identical to the tie-corrected τ_b of Table 2. Statistical resolution is not economic materiality, whether dependence changes a

decision quantity, which is a separate question for the capital experiment (Section 3.5); the experiment uses the calibrated latent coefficient, not the MPLE $\hat{\rho}$.

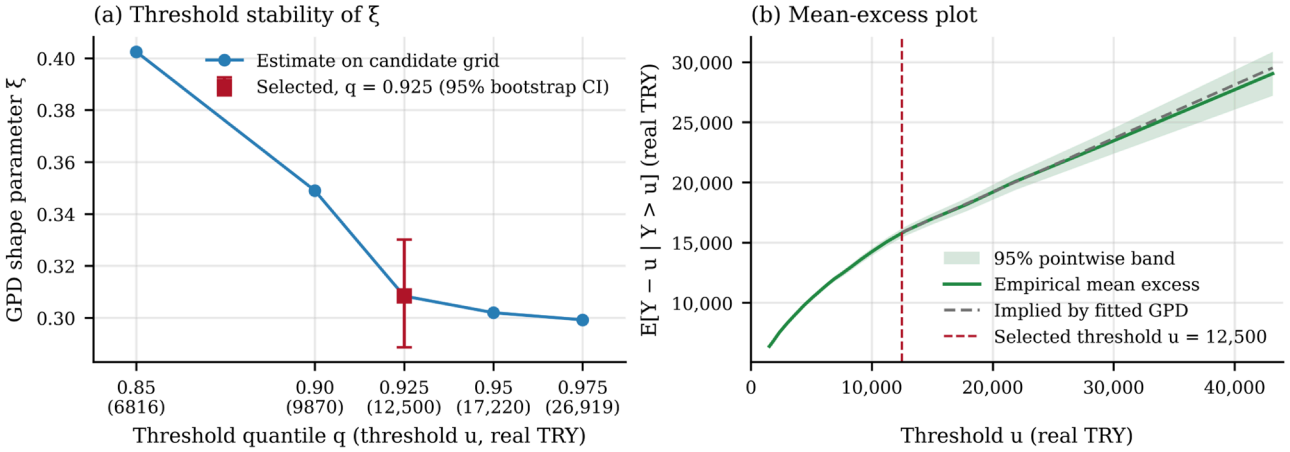

Figure 6. Generalized Pareto distribution (GPD) threshold diagnostics. (**a**) Shape parameter ξ (tail index $1/\xi$) across candidate threshold quantiles q (threshold u in real TRY in parentheses), with the exceedance-bootstrap 95% confidence interval at the selected threshold ($q = 0.925$, u = TRY 12,500). (**b**) Empirical mean excess $E[Y - u|Y > u]$ of claim severity Y with a 95% pointwise band, and the fitted-GPD mean excess $\{\beta + \xi(u - u^*)\}/(1 - \xi)$ above the selected threshold u^* (dashed line); β, GPD scale.

Table 4. Best-fitting copulas by axis. Candidates: independence, Joe, Frank, Gumbel, Clayton, survival Clayton, and Gaussian copulas, fitted by maximum pseudo-likelihood. ΔAIC and ΔBIC are relative to the independence copula (negative values favor the fitted family). 95% CIs: policy-cluster bootstrap percentile intervals ($B = 500$; jitter redrawn and model refitted in each resample). Implied τ: Kendall's τ of the fitted copula ($(2/\pi)arcsin\rho$ for the Gaussian family). AIC/BIC, Akaike/Bayesian information criterion; CI, confidence interval; λ_L, λ_U, lower/upper tail-dependence coefficients.

Axis	Best Family	Estimate (95% CI)	Implied τ (95% CI)	ΔAIC	ΔBIC	Tail Dependence
Frequency–severity	Gaussian	$\hat{\rho} = -0.0177$ [−0.0196,−0.0100]	−0.0113 [−0.0125,−0.0063]	−65	−55	$\lambda_L = \lambda_U = 0$
Settlement lag	Joe	$\hat{\theta}_J = 1.0461$ [1.0436,1.0504]	+0.0259 [+0.0246,+0.0283]	−1173	−1163	$\lambda_U = 0.060$

Table 5. Ledger of dependence parameters. Calibrated and prescribed Kendall coefficients enter the Gaussian generators through $\rho = sin(\pi\tau/2)$. Copula parameters: MPLE on jittered PITs, with policy-cluster bootstrap 95% CIs ($B = 500$). The naive and lag coefficients never enter the capital simulator. PIT, probability integral transform.

Quantity	Value [95% CI]	Origin; Use
Observed $\hat{\tau}_{FS}$ (τ_b)	−0.0065 [−0.0098,−0.0035]	Table 2; calibration target, equivalence test
Gaussian copula $\hat{\rho}$ (FS)	−0.0177 [−0.0196,−0.0100]	Table 4; selected by AIC and BIC; not a capital input
Joe copula $\hat{\theta}_J$ (lag)	1.0461 [1.0436,1.0504]	Table 4; selected by AIC and BIC; reserving axis only
Latent τ^*_{lat} (ρ^*)	−0.0102 (−0.0160)	calibrated to $\hat{\tau}_{FS}$ (slope $c = 0.644$); main scenario
Empirical severity ranks	simulated $\tau_b = -0.0070$	PITs by count class 1, 2, 3, 4+; copula-free check
Sweep τ_{lat}	{±0.05,±0.10,±0.20,±0.40}	prescribed; claim-level τ_b from −0.27 to +0.27
Artifact-experiment τ	{0,0.05,0.10,0.20}	prescribed (Section 2.3)
Positive-control τ	{0,0.10,0.25,0.40}	prescribed (Section 2.5)

3.5. Consequences for the Solvency-II-Style Capital Proxy

Table 6 and Figure 7 report the calibrated paired simulation of Section 2.7 ($\rho^* = -0.0160$). Assuming independence changes the 99.5% VaR by +0.008% (95% Monte Carlo interval [−0.034%,+0.030%]) and the capital distance K (TRY 4.83 million) by +0.18%

([−0.80%,+0.71%]), about TRY 9000 (first-order analytic value +0.14%). No VaR or TVaR level moves by more than 0.013%, the expected aggregate losses of the two arms agree to within 0.0003% (TRY 110.57 million), as the common claim-level margin implies, and no single seed block drives the result: the four per-block estimates of the VaR change range from −0.035% to +0.009% (Section S3).

Table 6. Paired capital results under the calibrated count-coupled Gaussian link ($\rho^* = -0.0160$; 100,000 policies; 20,000 paired replications): percentage change of the independent relative to the dependent arm (positive: independence gives the higher figure), with paired-bootstrap 95% Monte Carlo intervals ($B = 1000$). $K_q = VaR_q(S) - E(S)$; $q = 99.5\%$ defines the capital proxy.

Level q	VaR Change (%)	TVaR Change (%)	K_q Change (%)
95.0%	+0.013 [−0.001,+0.018]	+0.005 [+0.003,+0.007]	+0.49 [−0.04,+0.68]
99.0%	−0.002 [−0.019,+0.030]	+0.005 [−0.003,+0.012]	−0.04 [−0.50,+0.78]
99.5% (capital proxy)	+0.008 [−0.034,+0.030]	+0.008 [−0.006,+0.019]	+0.18 [−0.80,+0.71]

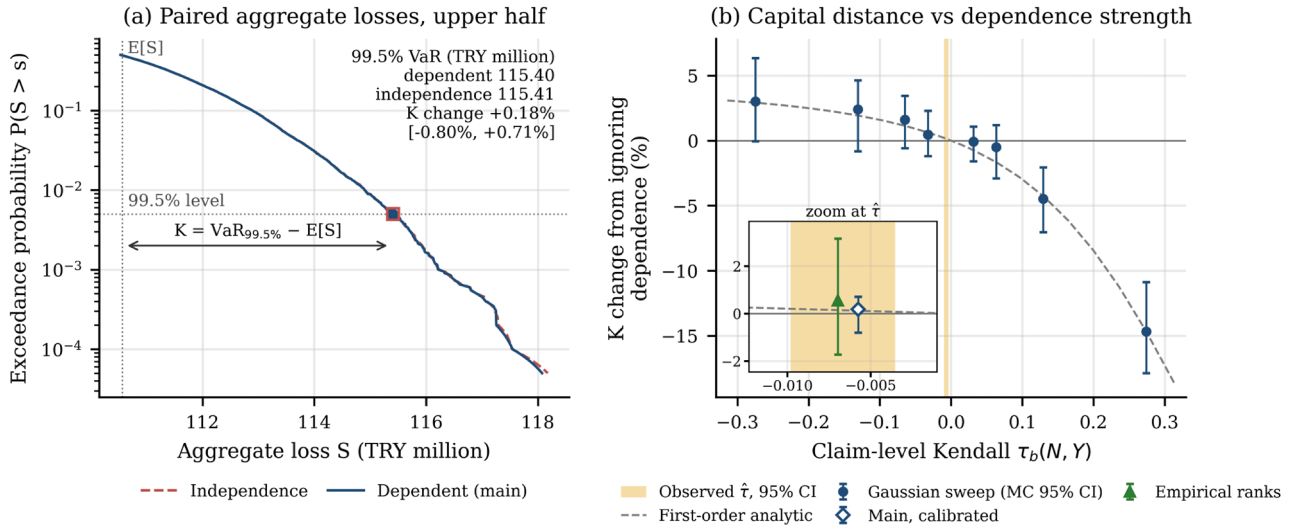

Figure 7. Capital effect of the observed association. (**a**) Upper half of both arms' aggregate-loss distributions, which overlap. (**b**) Change in K from ignoring dependence against claim-level $\tau_b(N,Y)$: Gaussian sweep (95% Monte Carlo intervals), first-order analytic curve; band, observed $\hat{\tau}$ (95% CI); inset, main and empirical-rank scenarios.

The copula-free empirical-rank scenario gives +0.021% ([−0.076%,+0.129%]) for the 99.5% VaR and +0.56% ([−1.73%,+3.15%]) for K (first-order value +1.1%). Its effect is larger because the empirical ranks also reproduce the observed fall in mean severity from one-claim to two-claim policies (TRY 4593 to 3711) and the lower dispersion of claims on multi-claim policies, which the Gaussian link implies only weakly (a 2% decline; Section S3). The analytic check agrees with the empirical severity moments by count class; $\mathrm{Var}(S_i)$ is 2.3% below its independence value (ratio 0.977), so the capital distance is, to first order, about 1.2% lower under the observed association; ignoring that association is therefore mildly conservative and immaterial.

In the sweep (Figure 7b), ignoring a negative association raises K by at most 3.0% down to $\tau_b = -0.27$, whereas ignoring a positive one understates K by 4.5% ([2.1%,7.0%]) at $\tau_b = 0.13$ and by 14.7% ([10.9%,17.9%]) at $\tau_b = 0.27$; at the latter point the 99.5% VaR itself moves by only 0.72%, because the expected loss dominates the aggregate VaR, which is why materiality is judged on the capital distance. By the analytic curve, under the calibrated Gaussian link and the fitted margins of this portfolio, the 5% reference level is reached only at $\tau_b \approx 0.14$, about twenty times the observed magnitude and well beyond the equivalence margin $\Delta = 0.02$, and a 1% understatement at $\tau_b \approx 0.04$. The split-sample Kupiec check

counts 48 exceedances of the simulated 99.5% quantile against 50 expected ($p = 0.775$; exact $p = 0.840$).

Whereas strong injected dependence ($\tau = 0.40$) moved a policy's 99.5% loss quantile by 11.5% with the claim-severity margin held fixed (Section 3.3), the observed association leaves the capital proxy of this well-diversified book statistically indistinguishable from independence and economically immaterial.

3.6. Consequences for Premium Pricing: Discrimination and Calibration

Since the capital proxy is practically unchanged, pricing can use independent frequency and severity margins, and the question becomes where modeling effort pays off. On the test years (Table 7; Figure 8), with all eight models fitted on the same 356,846 policies, most of the gain comes from modeling the pure premium rather than frequency alone: the frequency GLMs reach Gini values of 0.18, every pure-premium model exceeds 0.24, and the model selected on the validation year, the two-stage GBM (validation Gini 0.287), attains 0.255, 42% above the Poisson GLM (0.180); of the Gini gap of 0.076 between the two, the direct Tweedie GLM already closes all but 0.009. The test-year ranking is nearly flat at the top: the direct Tweedie GBM (0.256; $\Delta = +0.0012$ [−0.0015,+0.0038]), LightGBM (0.254), the two-stage GLM (0.253; $\Delta = -0.0023$ [−0.0065,+0.0019]), and XGBoost (0.252) are statistically indistinguishable from the selected model, so no conclusion depends on the selection; the direct Tweedie GLM trails it by 0.009, 4% of its Gini, whereas the frequency-only GLMs trail by 0.075. The validation-year Gini values (0.27–0.29 for the pure-premium models) exceed the test-year values because 2016 also served for early stopping and, for the Tweedie GLM, for the choice of p. The 2×2 factorial separates structure from learner. At fixed structure, the learner effect is +0.0102 ([+0.0069,+0.0136]) for direct and +0.0023 ([−0.0019,+0.0065]) for two-stage models; at fixed learner, the structure effect is +0.0067 ([+0.0050,+0.0085]) for GLMs and −0.0012 ([−0.0038,+0.0015]) for GBMs; the interaction is −0.0079 ([−0.0109,−0.0051]). Starting from the direct Tweedie GLM, boosting or a two-stage structure each closes most of the gap to the best model, but combining the two adds nothing. Refitting the GLMs on 2014–2016 changes their Gini values by at most 0.0025 and leaves these conclusions unchanged (Section S4).

Table 7. Pricing models on the test years 2017–2018, ordered by the test-year pure-premium Gini index, with paired-bootstrap 95% percentile confidence intervals (CIs; $B = 500$). All models are fitted on the same 356,846 policies of 2014–2015; *: highest validation-year (2016) Gini, carried forward. Δ: test-year Gini difference from that model; "Signif.": its CI excludes zero. GBM, gradient-boosted machine; GLM, generalized linear model.

Model	Valid. Gini	Test Gini	95% CI	Δ	95% CI for Δ	Signif.
Tweedie GBM	0.284	0.256	[0.245,0.269]	+0.0012	[−0.0015,+0.0038]	no
Two-stage GBM (Poisson × gamma) *	0.287	0.255	[0.244,0.267]	—	—	—
LightGBM (Tweedie)	0.283	0.254	[0.243,0.266]	−0.0011	[−0.0041,+0.0019]	no
Two-stage GLM (Poisson × gamma)	0.270	0.253	[0.241,0.266]	−0.0023	[−0.0065,+0.0019]	no
XGBoost (Tweedie)	0.283	0.252	[0.241,0.264]	−0.0029	[−0.0061,+0.0002]	no
Tweedie GLM ($p = 1.6$)	0.266	0.246	[0.234,0.259]	−0.0090	[−0.0133,−0.0047]	yes
Negative binomial GLM	0.184	0.181	[0.169,0.194]	−0.0740	[−0.0840,−0.0638]	yes
Poisson GLM	0.182	0.180	[0.168,0.192]	−0.0755	[−0.0855,−0.0652]	yes

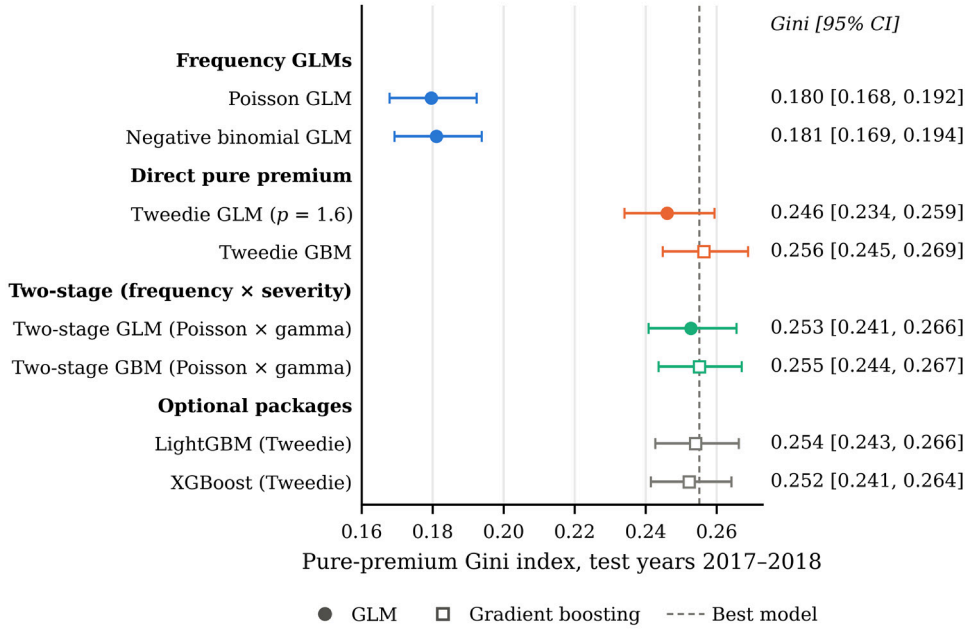

Figure 8. Test-year (2017–2018) pure-premium Gini index of every model in Table 7, grouped into frequency GLMs, direct pure-premium models, two-stage (frequency × severity) models, and optional boosting packages. Points: full-sample Gini; bars: paired-bootstrap 95% percentile intervals ($B = 500$); filled circles: GLMs; open squares: gradient boosting; dashed line: the model selected on the validation year (two-stage GBM). Colors distinguish the four model groups. Values are printed on the right; higher values indicate better discrimination. GBM, gradient-boosted machine; GLM, generalized linear model.

The lift (Figure 9b), which rises from 0.5 in the lowest to 2.2 in the highest decile of predicted pure premium, displays the risk differentiation that the Gini index summarizes. Calibration (Figure 9a) is assessed separately, and here the strict temporal design matters. At the level of its fitting years, the selected model over-predicts the test-year losses by about one sixth (predicted-to-observed ratio 1.167; decile ratios 1.01–1.25), whereas its decile-level slope, 0.96 ([0.92,1.00]), is compatible with one and the policy-level slope is 0.89 ([0.82,0.97]): the relative ordering and spread of the predictions are calibrated, but their level is not. The level shift has an identifiable source: claim frequency per policy-year falls by 6.9% a year over 2014–2016 (0.294, 0.275, 0.256; Table 1) and continues to fall in the test years, and the model, fitted on 2014–2015 without a trend term, carries the 2014–2015 level. Isotonic recalibration learned on 2016 removes only part of the excess (ratio 1.124) and leaves the Gini essentially unchanged (0.2551 $\rightarrow$ 0.2537) but worsens the slopes (decile slope 0.86), because the drift continues beyond the recalibration year; auto-calibration is guaranteed only on the fitting sample [39]. The on-level adjustment of Section 2.8 addresses the level directly (Table 8): re-leveling to 2016 alone brings the ratio to 1.113, the frequency trend from the fit-year center to the test-year center to 0.955, and their standard combination (factor 0.863) to 1.007, with decile ratios between 0.87 and 1.08 and the mean absolute log decile ratio reduced from 0.155 to 0.054; the Gini is unchanged by construction. The risk differentiation, the object of the Gini comparison, thus transfers across years, whereas the level requires the trend-based re-leveling that pricing practice applies each year; the combined factor is carried into the premium target of Section 3.7. The log–log slope on claimant policies only (0.58) conditions on a positive loss and is not a calibration slope (Section S4).

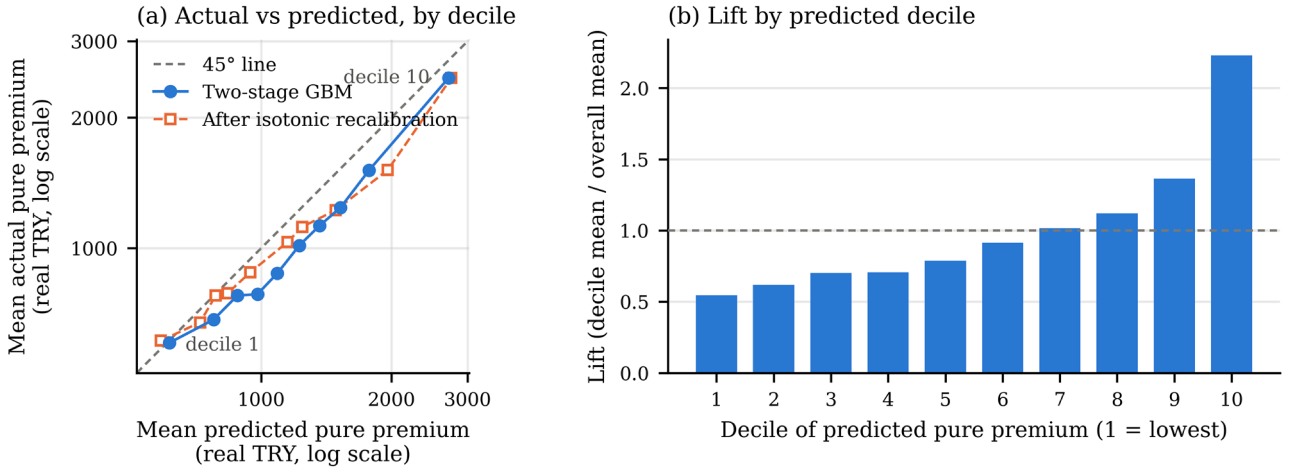

Figure 9. Calibration and lift of the selected model (two-stage GBM) on the test years 2017–2018, by decile of predicted pure premium. (**a**) Mean actual versus mean predicted pure premium (log–log axes) before and after isotonic recalibration fitted on 2016 (at the 2014–2015 level; the on-level factors of Table 8 rescale the horizontal axis); dashed: 45° line. (**b**) Decile mean actual pure premium relative to the overall test mean. GBM, gradient-boosted machine.

Table 8. On-level adjustment of the selected model (two-stage GBM) on the test years 2017–2018. Factors are estimated without test-year information: validation level, realized/predicted pure premium on 2016; frequency trend, log-linear trend of the claim frequency per policy-year over 2014–2016 (−6.9% per year), from the exposure-weighted center of the fit years (2014.55) to that of the test years (2017.44); combined, validation level followed by the trend from 2016 to the test-year center. P/O, predicted-to-observed ratio; the Gini index is invariant to a scalar factor.

Adjustment	Factor	Overall P/O	Decile P/O Range	Mean |Log Decile P/O|
None (2014–2015 level)	1.000	1.167	1.01–1.25	0.155
Validation level (2016)	0.953	1.113	0.97–1.19	0.114
Frequency trend, fit center → test center	0.818	0.955	0.83–1.02	0.055
Validation level, then trend 2016 → test center (carried forward)	0.863	1.007	0.87–1.08	0.054

3.7. RORAC-Target Technical Premium and Premium-Principle Differentiation

With the on-leveled expected losses, the capital proxy of the test portfolio, computed from the independent-margins aggregate, is TRY 7.85 million, 2.37% of the expected loss of TRY 330.4 million (aggregate skewness 0.038), 10% above a direct compound-Poisson Monte Carlo check (TRY 7.10 million; normal-approximation standard error TRY 0.33 million), so the Cornish–Fisher proxy is mildly conservative. At h = 12% the RORAC-target premium is TRY 441.8 million, with an expected loss ratio of 0.748 and a realized loss ratio of 0.743, the two agreeing once the level is on-leveled (at the 2014–2015 level the realized ratio would have been 0.64); the capital margin $h \cdot$ SCR is only TRY 0.94 million, 0.21% of premium (0.22% under the negative binomial compound variance): expected loss and expenses dominate. Premium principles therefore differ only in shape: by Remark 2, the expected-value, variance, exponential, and cost-of-capital principles induce identical risk-proportional loadings, whereas the standard-deviation principle is mildly regressive (Section S4). In relative decile adequacy (Text S1, Figure S1), both premiums stay within about 15% of risk-fair pricing: the model premium is relatively under-priced in the lowest-risk decile and the highest (relative loss ratios 1.15 and 1.06), close to risk-fair in deciles 2–3 (1.03), and over-priced in deciles 4–9 (0.93–0.99), whereas the market premium is flatter (0.92–1.07), its largest relative over-pricing falling in the highest-risk decile. Because relative loss ratios are normalized by the portfolio ratio of the same premium, they are invariant to the level adjustment.

3.8. Reserve Basis Reconciliation and Confidence-Level Choice

On the settlement-lag axis, larger claims settle slightly later ($\hat{\tau}_{\text{lag}} = +0.0185$), but insofar as this pattern is stable across accident years, the development factors absorb it; what is not absorbed is the choice of basis. Table 9 reconciles the three bases of Section 2.10. The paid chain-ladder reserve is TRY 689,034 (Mack standard error 541,623): the paid pattern is almost complete after the first development year (factors 1.116, 1.003, 1.001, and at most 1.001 thereafter), so the method learns that little is paid late. The case reserves open at the valuation date amount to TRY 9.44 million on 201 files, 105 of which, carrying TRY 5.4 million, belong to accident years 2014–2017 and are three or more years old. The incurred chain-ladder, whose factors are all at or below one (0.917 at the first development year, then 0.998, 0.999, 0.998, 1.000, and 0.999), estimates an outstanding liability of TRY 7.32 million (Mack standard error 560,914), TRY 2.1 million below the case reserves: case reserves in this book are systematically redundant. The historical run-off quantifies the redundancy. Over the mature December snapshots (2014–2017), 51.7% of the case reserves held on then-open files were subsequently paid (57.4% including the reserves still open at the valuation date), with 16% of the files paid nothing at all; by age of the file at the snapshot, the paid share is 53% under one year, 35% at one year, and 34% at two years or more (88% including still-open reserves). Applying the age-matched ratios to the files open at the valuation date gives an expected payment of TRY 3.2 million (4.9 million with the pooled ratio).

Table 9. Reserve bases at the valuation date (December 2020; real TRY at December-2018 prices). Paid and incurred chain-ladder reserves with Mack standard errors; case reserves on the 201 open claim files; run-off-adjusted expectation from the age-matched paid shares of the mature December snapshots (2014–2017; pooled ratio in parentheses); technical provision (TP) = best estimate (BE) plus the 6% cost-of-capital risk margin of Equation (6).

Basis	Amount (TRY)	Reading
Paid chain-ladder reserve	689,034 (s.e. 541,623)	paid pattern almost complete after development year 1; lowest of the bases
Case reserves on open files	9,435,710 (201 files)	105 files, TRY 5.4 million, from accident years 2014–2017; highest of the bases
Incurred chain-ladder outstanding	7,322,842 (s.e. 560,914)	incurred factors $\leq$1: case reserves systematically redundant
Run-off-adjusted expectation	3,191,722 (4,881,678)	51.7% of held case reserves paid in mature snapshots; 53%/35%/34% by age
TP, incurred basis	7,458,193 (1.02 $\times$ BE)	paid basis: 908,615 (1.32 $\times$ BE)

The discrepancy that motivated this section is therefore not an error of either method but a property of the book: its open files are dominated by a few old, large, disputed claims whose outcome the paid pattern, learned from an almost fully developed first year, underweights, and whose case reserves, learned from the insurer's conservative reserving, overstate. In this portfolio the paid chain-ladder gives the lowest, and the case reserves the highest of the bases, and the run-off-adjusted (TRY 3.2–4.9 million) and incurred (TRY 7.3 million) bases lie between them; these are material differences between reserving bases, not formal lower and upper bounds for the ultimate liability, which only future run-off can establish. The confidence-level choice is then made on the incurred basis, whose predictive uncertainty is small relative to its level (coefficient of variation 7.7%, against 78.6% on the paid basis): the 75% and 99.5% VaR reserves are 1.05 and 1.21 times the incurred best estimate, and the technical provision with the cost-of-capital risk margin of Equation (6) (TRY 135,351) is TRY 7.46 million, 1.02 times the best estimate, against

1.32 times on the paid basis (TRY 908,615), where the same margin is large relative to a small reserve. The full set of confidence-level reserves, the newsvendor-optimal reserves for penalty ratios from 1 to 10, and the distributional sensitivity on both bases are given in Text S1, Section S5; on the incurred basis, the four moment-matched predictive distributions differ by at most 1.3% on the decision-relevant penalty grid, so the choice of confidence level is, once the basis is settled, robust to the predictive distribution.

4. Discussion

The central contribution is a principle of order: measurement should precede model complexity. The count–average construction on which copula models of summarized claims are built detects association built into it, not learned from the data: in this portfolio, exact independence reproduces the whole of the observed count–average coefficient, and the residual has the opposite sign (Section 3.2). Clean measurement turns a strong positive naive association into a small negative one inside the equivalence region; under the Gaussian link and the margins of this portfolio, only a positive association about twenty times as large would change capital materially (Section 3.5). As the same workflow detects injected dependence in every replication (Section 3.3), the small estimate reflects the data, not the instrument.

Our negative covariate-adjusted count effect shares the sign found for collision and comprehensive coverages [8,9]; positive associations for motor liability [10] and aggregated count–average pairs [11,12] concern other coverages or aggregation levels. Oh et al. [15] establish that the count–average pair is dependent even under independence and model that structure; the decomposition of Section 3.2 adds the sign and size of the induced association on the rank scale, shows that it is driven by skewness rather than by the variance scaling of the average, and finds that in a real portfolio it can exceed the observed coefficient, so that the sign and strength of a copula fitted to this pair are not evidence about frequency–severity dependence. Careful modeling likewise leaves little apparent correlation between lines of business [14]. Bonus hunger, which can induce frequency–severity dependence in reported claims even when ground-up frequency and severity are independent [16], is a behavioral counterpart of the construction-induced associations studied here: an association measured in claims data need not belong to the loss process. What is new is the quantitative separation of measured from construction-induced association in the data at hand, with an equivalence test and a capital translation that state the size and cost of any dependence before it is modeled.

Modeling effort can go where it changes decisions. The entire capital loading is a fraction of a percent of premium (Section 3.7), whereas modeling the pure premium, more than the choice of learner, changes risk discrimination substantially; the pricing comparison is also the one place where modeling frequency and severity jointly rather than separately could pay off, and the small structure effect found there is the economic reading of the null dependence result. Calibration needs more than one diagnostic, since a level drift between fitting and test years can coexist with calibrated relative differentiation, and the drift is removed by a trend-based on-level adjustment, not by recalibration (Section 3.6). In reserving, the paid and incurred bases of this book differ by an order of magnitude, and reconciling them through the historical run-off of case reserves precedes any choice of confidence level (Section 3.8). The reserving analysis bears on the research question only through the settlement-lag axis and the choice of basis: the weak lag association is absorbed by the development factors, and the basis, not a dependence margin, governs the reserve level; it establishes no further dependence result.

5. Conclusions and Recommendations

5.1. Conclusions

This paper turned a seemingly negative result, a small frequency–severity association in a motor casco portfolio, into a methodological program: the construction-induced association of the count–average pair was given its mechanism on the rank scale and decomposed in the data; a leakage-safe protocol was validated by a replicated positive control; and an equivalence test and a paired capital experiment showed that the observed association leaves the capital proxy practically unchanged. Three findings answer the research question. First, artifact-safe measurement reverses the naive verdict: the observed count–average association is reproduced in full under exact independence, and what remains is a small negative association inside the equivalence region, whereas injected dependence is detected in every replication. Second, propagating the observed association through a calibrated, leakage-free simulator changes the 99.5% capital proxy by an amount indistinguishable from Monte Carlo noise, so a dependence layer is not warranted for this portfolio. Third, in the secondary decision layer, modeling the pure premium rather than frequency alone drives risk differentiation, a trend-based on-level adjustment restores the premium level, and the paid, incurred, and case-reserve bases must be reconciled before a reserve confidence level is chosen.

5.2. Recommendations for Practice

For practitioners and model validators, the protocol reduces to three checks and one allocation rule. First, measure dependence at claim level rather than count against total or average, with a cluster-robust interval and the minimum detectable effect, so that a null is not a power failure; where only aggregates exist, the count–average pair may still be modeled jointly, but its association should be decomposed against a permutation null before it is read as frequency–severity dependence. Second, validate the measurement chain on synthetic portfolios with injected dependence, and calibrate simulators to the observed claim-level coefficient, which understates the latent dependence. Third, test the estimate for equivalence against a stated margin and translate it into the model's economic metric, judged against a stated materiality level, before adding a dependence layer. Our recommendation is therefore selective rather than dismissive: copula modeling is warranted once the dependence signal has survived clean measurement, a power assessment, and validation. Fourth, once dependence is settled, spend the modeling effort where it changes decisions: on the pure-premium structure and its trend-based re-leveling in pricing, and on reconciling the paid, incurred, and case-reserve bases before a reserve confidence level is set; the cost-based choice of that level (Text S1, Section S5) is secondary to the basis.

5.3. Limitations and Future Research

1. Scope: the evidence comes from one insurer's short-tail, attritional motor own-damage portfolio; in long-tail or catastrophe-exposed lines, where settlement dynamics and common shocks are stronger, dependence and its cost may be materially larger, although careful modeling often reduces apparent correlations between lines of business to zero-to-modest levels [14], so any such dependence should itself be measured with artifact-safe diagnostics. The research question is scoped to the measurement problem and to this class of portfolios; the economic conclusions need not transfer, and the positive control supports internal validity only.
2. Ties: heavily tied counts and delays weaken copula-based tests, so the Vuong and Cramér–von Mises tests are secondary evidence.
3. Premiums: the scope of the recorded premium field cannot be verified, so only relative adequacy is interpreted.

4. Reserving: the paid chain-ladder understates the outstanding liability of a book whose open files are dominated by a few old, large, disputed claims, and the case reserves overstate it; the incurred and run-off-adjusted bases lie between them (Section 3.8); none of the bases is a formal bound for the ultimate liability, which only future run-off can establish. The analysis reconciles the bases and illustrates the choice of confidence level; it does not assess the adequacy of the insurer's booked reserves.
5. Allocation: in a compound-Poisson model with a common claim-size distribution, variance-proportional shares equal the Euler shares exactly (Remark 2); only the Cornish–Fisher capital total is approximate (skewness correction about 1.4% of capital; 10% above a direct Monte Carlo check), and the whole capital margin is about 0.21% of premium, so no plausible allocation error would change an individual premium materially.
6. Estimands: the regression asks whether, among comparable policies, claims on multi-claim policies are smaller (they are somewhat smaller); the capital experiment asks whether the observed claim-level association changes capital in a simulator that reproduces observed claim counts and sizes.
7. Materiality: the 5% change in the capital distance is a convention of this study, not a regulatory threshold; the association at which it is reached (Section 3.5) is specific to the Gaussian link and the margins of this portfolio, and a different reference level or a long-tail line could move the verdict.

Future work includes long-tail lines, where dependence is plausibly stronger, the decomposition of count–average associations in other portfolios and lines, and dynamic or covariate-conditional dependence.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math14214104/s1, Text S1: Supplementary methods (Sections S1–S6: Statements and proofs of the count–sum property and the shared-uniform leakage result, with the size of all three artifacts in the controlled experiment, and details on Remark 1; secondary independence tests and the equivalence-margin grid; capital-experiment implementation details; pricing-model details, including premium principles and relative adequacy; reserving details, including confidence-level and newsvendor reserves on the paid and incurred bases; independent re-derivation of the headline results); Data S1: Aggregate, anonymized derived data underlying all tables and figures, including the verification table, with a README mapping each file to its source (no policyholder-level or personal data). Text S1 also contains Figure S1: Relative risk-decile premium adequacy on the test years 2017–2018; Figure S2: Predictive distribution of the total reserve on the paid basis; Figure S3: Newsvendor-optimal reserve confidence level against the penalty ratio; Table S1: Decile predicted-to-observed ratios of the selected model under the on-level factors; Table S2: Reserve strategies on the paid basis; Table S3: Distributional sensitivity of the reserve on the paid basis; Table S4: Reserve strategies on the incurred basis; Table S5: Independent re-derivation of the headline results. References [6,7,9,11,17,23–25,30–33,36,39,40,42–47] are cited in the Supplementary Materials.

Author Contributions: Conceptualization, Y.A. and F.K.; methodology, Y.A.; software, Y.A.; validation, Y.A. and F.K.; formal analysis, Y.A.; investigation, Y.A.; data curation, Y.A.; writing—original draft preparation, Y.A.; writing—review and editing, F.K.; visualization, Y.A.; supervision, F.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors are grateful to the insurer that supplied the data. Views expressed here are the authors' own and do not necessarily represent the positions of Türk Reasürans. During the preparation of this manuscript, the authors used Claude Fable 5.1 (Anthropic) and ChatGPT GPT-5 (OpenAI) for the purposes of drafting and language editing, assistance in developing and checking the analysis and simulation code, and discussion and verification support for the formal results. The authors have reviewed and edited the output and take full responsibility for the content of this publication. All analyses and simulations were conducted by the authors, and all numerical results were obtained from the authors' own executions of the code.

Conflicts of Interest: Author Y.A. is employed by Türk Reasürans. Türk Reasürans had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results. Author F.K. declares no conflicts of interest.

Appendix A. Full Covariate-Adjusted Severity Regression

Table A1 reports the full coefficient vector of the covariate-adjusted log-severity regression (2) of Section 2.4 (n = 214,283 claims on 171,822 policies, which form the clusters of the bootstrap; R^2 = 0.028). Inference uses the policy-cluster bootstrap of Section 2.4, which resamples policies with all their claims: the reported standard errors are the standard deviations of the B = 500 bootstrap estimates, and the confidence intervals are percentile intervals. The count coefficient is the estimate carried to Table 2 and Section 3.2. Log exposure enters because the claim count depends on exposure; its inclusion leaves the count coefficient essentially unchanged, consistent with the predominance of annual policies.

Variance-inflation factors in the full design are 1.01 (claim count), 1.15 (log exposure), 1.44 (vehicle age), 3.41 (insured age), 3.74 (horsepower), 3.05 (displacement), and 2.86 (vehicle value); they are 1.9–2.6 for the accident-year dummies and 3.59 for the corporate-policyholder dummy. The insured-age and corporate values reflect the coding of corporate policies with insured age zero, which the corporate dummy absorbs. Pairwise correlations among vehicle value, horsepower, and displacement range from 0.56 to 0.79 in the claim-level sample, so the engineering block is moderately collinear but well below conventional concern thresholds, and the claim-count coefficient is essentially orthogonal to it.

Two terms rest on very few observations. The log-exposure coefficient has a wide, asymmetric interval because five claims on policies with zero recorded exposure enter at the floor $\log(10^{-6})$ of the logarithm. Body type is retained to keep the rating vector complete, but its dummy is identified by a single claim on one policy, since almost all policies share the modal body type (Section 2.1); its estimate is not interpretable and is suppressed in Table A1. Neither term affects the count coefficient: dropping the five zero-exposure claims, or dropping the body-type dummy, leaves β_N at -0.0436.

Table A1. Covariate-adjusted log-severity regression: coefficients with policy-cluster bootstrap standard errors (SE) and 95% percentile confidence intervals (CI) (n = 214,283 claims; R^2 = 0.028; B = 500). Reference categories: accident year 2014, male policyholder, modal body type. std., standardized covariate.

Term	Estimate	Cluster SE	95% CI
Intercept	7.304	0.012	[7.282,7.328]
Claim count N	−0.0436	0.0040	[−0.0515,−0.0355]
Accident year 2015	−0.002	0.012	[−0.027,+0.020]
Accident year 2016	+0.061	0.012	[+0.036,+0.082]
Accident year 2017	+0.145	0.012	[+0.119,+0.168]
Accident year 2018	+0.092	0.013	[+0.066,+0.117]
Accident year 2019	+0.017	0.021	[−0.020,+0.061]

Table A1. *Cont.*

Term	Estimate	Cluster SE	95% CI
log(earned exposure) [1]	+0.042	0.097	[−0.073,+0.297]
Vehicle age (std.)	−0.012	0.004	[−0.020,−0.005]
Insured age (std.)	−0.006	0.006	[−0.017,+0.005]
Horsepower (std.)	+0.117	0.007	[+0.104,+0.131]
Displacement (std.)	−0.011	0.006	[−0.023,+0.002]
Vehicle value (std.; nominal)	+0.140	0.006	[+0.128,+0.151]
Gender: female	−0.005	0.008	[−0.022,+0.010]
Gender: corporate	−0.027	0.015	[−0.054,+0.002]
Body type (non-modal) [2]	—	—	—

[1] Interval widened by claims on policies with zero recorded exposure (see text). [2] Not interpretable; estimate suppressed (see text).

Appendix B. Specification of the Pricing Models

Table A2 gives the full specification of the pricing models of Section 2.8. All models use the same inputs—the five numeric rating variables (vehicle age, insured age, horsepower, engine displacement, and insured vehicle value) and dummy-coded gender and body type—and are evaluated on the 295,853 policies of the test years 2017–2018. The numeric variables are standardized with the means and standard deviations of the 2014–2015 fit years, applied unchanged to the validation and test years. All models, GLMs and boosted models alike, are fitted on the same 356,846 policies of 2014–2015; the severity stage of both two-stage models uses all 79,760 claimant policies of these years with a positive claim cost. The negative binomial GLM has variance $\mathrm{Var}(N|x) = \mu(x) + \alpha\,\mu(x)^2$ with α estimated by the method of moments ($\hat{\alpha} = 0.87$). Only the number of trees (by early stopping on 2016) and the variance power of the Tweedie GLM (by profile likelihood on 2016; Section S4) were selected on the data; all other settings were fixed a priori rather than tuned. No boosted model uses row or column subsampling or an exposure offset, since earned exposure is close to one year for almost all policies (Table 1). LightGBM (version 4.6.0) and XGBoost (version 3.3.0) run single-threaded with a fixed seed and predict with the early-stopped model.

Table A2. Specification of the pricing models: panel (**a**) generalized linear models (GLMs), panel (**b**) gradient-boosted machines (GBMs). NB2, negative binomial with quadratic variance $\mu + \alpha\mu^2$; IRLS, iteratively reweighted least squares (Fisher scoring); ML, maximum likelihood; p, Tweedie variance power.

(a) GLMs	Poisson GLM	NB2 GLM	Two-Stage GLM	Tweedie GLM
Response and family	claim count; Poisson, log link	claim count; NB2, log link	count (Poisson) × mean severity (gamma, log link)	pure premium; Tweedie, log link
Offset and weights	log exposure	log exposure	log exposure (frequency); claim counts as weights (severity)	none
Fit years (policies)	2014–2015 (356,846)	2014–2015 (356,846)	2014–2015 (356,846; severity: 79,760 claimant policies with a positive amount)	2014–2015 (356,846), the same fits as the p-selection
Tuning in 2016	none (no tuning parameter)	none	none	p selected by profile log-likelihood
Dispersion or power	—	$\hat{\alpha} = 0.87$ (method of moments on the Poisson fit)	—	$\hat{p} = 1.60$ (grid 1.10–1.90, step 0.05; ϕ by ML)
Estimation	IRLS, ≤40 iterations	IRLS, ≤40 iterations	IRLS, ≤40 iterations	IRLS, ≤40 iterations

Table A2. *Cont.*

(b) GBMs	Tweedie GBM	Two-Stage GBM	LightGBM	XGBoost
Implementation	in-house histogram GBM (second-order leaf values)	in-house histogram GBM (second-order leaf values)	LightGBM 4.6.0	XGBoost 3.3.0 (histogram method)
Loss	Tweedie deviance, $p = 1.5$	Poisson deviance (frequency); gamma deviance, claim counts as weights (severity)	Tweedie, $p = 1.5$	Tweedie, $p = 1.5$
Fit data (2014–2015)	all 356,846 policies	frequency: all 356,846 policies; severity: all 79,760 claimant policies with a positive amount	all 356,846 policies	all 356,846 policies
Early stopping (2016)	Tweedie deviance, all policies; patience 15, at most 1000 trees	Poisson deviance, all policies; gamma deviance, 34,388 claimant policies; patience 15, at most 1000 trees	Tweedie negative log-likelihood, all policies; patience 20, at most 500 rounds	Tweedie negative log-likelihood, all policies; patience 20, at most 500 rounds
Selected trees	51	362 (frequency); 106 (severity)	61	80 (of 100 trained)
Learning rate	0.08	0.08	0.05	0.05
Tree size	depth 3	depth 3	31 leaves (depth unrestricted)	depth 5
Histogram bins	32 (quantiles of the fit years)	32 (quantiles of the fit years)	255 (default)	256 (default)
Minimum node size	50 observations to split; 20 per child	50 observations to split; 20 per child	100 observations per leaf	minimum child weight 100
L_2 leaf penalty	1	1	0 (default)	1 (default)
Subsampling	none	none	none	none
Threads and seed	deterministic (no subsampling)	deterministic (no subsampling)	1 thread, deterministic mode, seed 20240517	1 thread, seed 20240517

References

1. Lee, G.Y.; Shi, P. A Dependent Frequency–Severity Approach to Modeling Longitudinal Insurance Claims. *Insur. Math. Econ.* **2019**, *87*, 115–129. [CrossRef]
2. Oh, R.; Shi, P.; Ahn, J.Y. Bonus-Malus Premiums Under the Dependent Frequency–Severity Modeling. *Scand. Actuar. J.* **2020**, *2020*, 172–195. [CrossRef]
3. Sklar, A. Fonctions de répartition à *n* Dimensions Et Leurs Marges. *Publ. Inst. Stat. Univ. Paris* **1959**, *8*, 229–231.
4. Joe, H. *Dependence Modeling with Copulas*; CRC Press: Boca Raton, FL, USA, 2014.
5. Frees, E.W.; Valdez, E.A. Understanding Relationships Using Copulas. *N. Am. Actuar. J.* **1998**, *2*, 1–25. [CrossRef]
6. Denuit, M.; Dhaene, J.; Goovaerts, M.; Kaas, R. *Actuarial Theory for Dependent Risks: Measures, Orders and Models*; Wiley: Chichester, UK, 2005.
7. Genest, C.; Nešlehová, J. A Primer on Copulas for Count Data. *ASTIN Bull.* **2007**, *37*, 475–515. [CrossRef]
8. Gschlößl, S.; Czado, C. Spatial Modelling of Claim Frequency and Claim Size in Non-Life Insurance. *Scand. Actuar. J.* **2007**, *2007*, 202–225. [CrossRef]
9. Garrido, J.; Genest, C.; Schulz, J. Generalized Linear Models for Dependent Frequency and Severity of Insurance Claims. *Insur. Math. Econ.* **2016**, *70*, 205–215. [CrossRef]
10. Shi, P.; Feng, X.; Ivantsova, A. Dependent Frequency–Severity Modeling of Insurance Claims. *Insur. Math. Econ.* **2015**, *64*, 417–428. [CrossRef]
11. Czado, C.; Kastenmeier, R.; Brechmann, E.C.; Min, A. A Mixed Copula Model for Insurance Claims and Claim Sizes. *Scand. Actuar. J.* **2012**, *2012*, 278–305. [CrossRef]
12. Krämer, N.; Brechmann, E.C.; Silvestrini, D.; Czado, C. Total Loss Estimation Using Copula-Based Regression Models. *Insur. Math. Econ.* **2013**, *53*, 829–839. [CrossRef]
13. Embrechts, P.; McNeil, A.J.; Straumann, D. Correlation and Dependence in Risk Management: Properties and Pitfalls. In *Risk Management: Value at Risk and Beyond*; Dempster, M.A.H., Ed.; Cambridge University Press: Cambridge, UK, 2002; pp. 176–223.
14. Avanzi, B.; Taylor, G.; Wong, B. Correlations Between Insurance Lines of Business: An Illusion or a Real Phenomenon? Some Methodological Considerations. *ASTIN Bull.* **2016**, *46*, 225–263. [CrossRef]
15. Oh, R.; Ahn, J.Y.; Lee, W. On Copula-Based Collective Risk Models: From Elliptical Copulas to Vine Copulas. *Scand. Actuar. J.* **2021**, *2021*, 1–33. [CrossRef]

16. Park, S.C.; Kim, J.H.T.; Ahn, J.Y. Does Hunger for Bonuses Drive the Dependence Between Claim Frequency and Severity? *Insur. Math. Econ.* **2018**, *83*, 32–46. [CrossRef]
17. Mack, T. Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates. *ASTIN Bull.* **1993**, *23*, 213–225. [CrossRef]
18. Wüthrich, M.V.; Merz, M. *Stochastic Claims Reserving Methods in Insurance*; Wiley: Chichester, UK, 2008.
19. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. Erratum in *Nat Methods* **2020**, *17*, 352. https://doi.org/10.1038/s41592-020-0772-5. [CrossRef] [PubMed]
20. Kendall, M.G. The Treatment of Ties in Ranking Problems. *Biometrika* **1945**, *33*, 239–251. [CrossRef] [PubMed]
21. Field, C.A.; Welsh, A.H. Bootstrapping Clustered Data. *J. R. Stat. Soc. Ser. B* **2007**, *69*, 369–390. [CrossRef]
22. Efron, B. Better Bootstrap Confidence Intervals. *J. Am. Stat. Assoc.* **1987**, *82*, 171–185. [CrossRef]
23. Vuong, Q.H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **1989**, *57*, 307–333. [CrossRef]
24. Genest, C.; Rémillard, B. Tests of Independence and Randomness Based on the Empirical Copula Process. *Test* **2004**, *13*, 335–369. [CrossRef]
25. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
26. Schuirmann, D.J. A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680. [CrossRef] [PubMed]
27. Pickands, J., III. Statistical Inference Using Extreme Order Statistics. *Ann. Stat.* **1975**, *3*, 119–131. [CrossRef]
28. Embrechts, P.; Klüppelberg, C.; Mikosch, T. *Modelling Extremal Events for Insurance and Finance*; Springer: Berlin, Germany, 1997.
29. McNeil, A.J. Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory. *ASTIN Bull.* **1997**, *27*, 117–137. [CrossRef]
30. European Parliament and Council Directive 2009/138/EC on the Taking-up and Pursuit of the Business of Insurance and Reinsurance (Solvency II). *Off. J. Eur. Union* **2009**, *L335*, 1–155.
31. Kupiec, P.H. Techniques for Verifying the Accuracy of Risk Measurement Models. *J. Deriv.* **1995**, *3*, 73–84. [CrossRef]
32. Jørgensen, B. Exponential Dispersion Models. *J. R. Stat. Soc. Ser. B* **1987**, *49*, 127–145. [CrossRef]
33. Smyth, G.K.; Jørgensen, B. Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling. *ASTIN Bull.* **2002**, *32*, 143–157. [CrossRef]
34. Ohlsson, E.; Johansson, B. *Non-Life Insurance Pricing with Generalized Linear Models*; Springer: Berlin, Germany, 2010.
35. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
36. Dunn, P.K.; Smyth, G.K. Series Evaluation of Tweedie Exponential Dispersion Model Densities. *Stat. Comput.* **2005**, *15*, 267–280. [CrossRef]
37. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
38. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
39. Wüthrich, M.V.; Ziegel, J. Isotonic Recalibration Under a Low Signal-to-Noise Ratio. *Scand. Actuar. J.* **2024**, *2024*, 279–299. [CrossRef]
40. Cornish, E.A.; Fisher, R.A. Moments and Cumulants in the Specification of Distributions. *Rev. Inst. Int. Stat.* **1938**, *5*, 307–320. [CrossRef]
41. Tasche, D. Allocating Portfolio Economic Capital to Sub-Portfolios. In *Economic Capital: A Practitioner Guide*; Dev, A., Ed.; Risk Books: London, UK, 2004; pp. 275–302.
42. Bühlmann, H. *Mathematical Methods in Risk Theory*; Springer: Berlin, Germany, 1970.
43. Deprez, O.; Gerber, H.U. On Convex Principles of Premium Calculation. *Insur. Math. Econ.* **1985**, *4*, 179–189. [CrossRef]
44. England, P.D.; Verrall, R.J. Stochastic Claims Reserving in General Insurance. *Br. Actuar. J.* **2002**, *8*, 443–518. [CrossRef]
45. England, P.; Verrall, R. Analytic and Bootstrap Estimates of Prediction Errors in Claims Reserving. *Insur. Math. Econ.* **1999**, *25*, 281–293. [CrossRef]
46. *IFRS Foundation IFRS 17 Insurance Contracts*; IFRS Foundation: London, UK, 2017.
47. Arrow, K.J.; Harris, T.; Marschak, J. Optimal Inventory Policy. *Econometrica* **1951**, *19*, 250–272. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.