Predictive Modeling of Workers' Compensation Injury Frequency Using Federal OSHA Data

Fred Duggan

MarisRisk

March 2026

Technical White Paper

Table of Contents

1. Executive Summary

This study develops and validates a predictive model for workers' compensation injury frequency using 692,412 establishment-year observations from OSHA's Injury Tracking Application (ITA) spanning 2016-2024. The model uses **year Y** features (prior injury rates, enforcement history, industry classification, employer size) to predict **year Y+1** recordable injury counts.

> **Key Findings:**
>
> - The model achieves a **8.7x lift ratio** between the top and bottom risk deciles on an out-of-time test set (2022→2023), indicating strong discrimination.
> - The top risk decile captures **17.9% of all recordable injuries**, compared to 15.1% captured by the current NAICS-4 percentile peer grading approach.
> - The predictive signal reflects **persistent establishment-level risk differences**, not simply entity size: prior injury rate predicts future rate at Spearman ρ = 0.48, while entity size predicts rate at only ρ = 0.12.
> - Discriminatory power **attenuates for smaller entities** as expected (ρ = 0.31 for <10 FTE vs ρ = 0.78 for 250+ FTE). For micro-employers, predictions are heavily weighted toward

industry class rates, with limited individual experience differentiation.
- Buhlmann-Straub credibility weighting blends individual and group estimates: micro-entities (<5 FTE) receive 83% weight on the NAICS group rate, while large entities (>100 FTE) are predominantly experience-rated.

Important limitations: The OSHA ITA training population is not representative of all US employers. It is tilted toward larger establishments (250+ employees) and designated high-hazard industries. Generalization to small, low-hazard, or non-ITA-reporting employers is limited; predictions for such entities should be treated as class-rate priors with modest risk modifiers, not as individually calibrated estimates. The model also exhibits ~25% aggregate over-prediction on the 2022→2023 test set due to secular trend, requiring a calendar-year adjustment before production use. See Sections 7 and 12 for details.

The model architecture combines a Negative Binomial GLM with exposure offset (providing the interpretable base rate), Buhlmann-Straub credibility weighting (blending individual experience with industry priors), and a LightGBM model trained on GLM residuals (capturing nonlinear interactions, improving validation MAE by 19.1%). The GLM layer is fully transparent and standard in actuarial practice. The gradient boosting layer improves discrimination but introduces model governance considerations and explainability requirements discussed in Section 6. Model deployment is contingent on calendar-year adjustment and governance sign-off.

2. Data Description

2.1 Primary Data Source

The OSHA Injury Tracking Application (ITA) collects annual injury and illness data from establishments meeting reporting thresholds (generally 250+ employees, or 20+ employees in designated high-hazard industries). Each record represents one establishment-year and includes:

- **Exposure:** Total hours worked, annual average employee count
- **Outcomes:** Total recordable cases, days-away-from-work cases, job transfer/restriction cases, fatalities
- **Derived rates:** TRIR (Total Recordable Incident Rate = cases × 200,000 / hours), DART rate
- **Classification:** 6-digit NAICS code, EIN, establishment identifier

2.2 Supplementary Data Sources

- **OSHA Enforcement:** 5.1M inspections and 13.2M violations with severity classifications and penalty amounts
- **OSHA Severe Injury Reports (SIR):** 103K events including hospitalizations, amputations, fatalities

2.3 Panel Construction

The model uses a **lagged panel design**: for each establishment with consecutive-year ITA records, we pair year Y features with year Y+1 outcomes. This produces 692,412 observation pairs across 267,091 unique establishments.

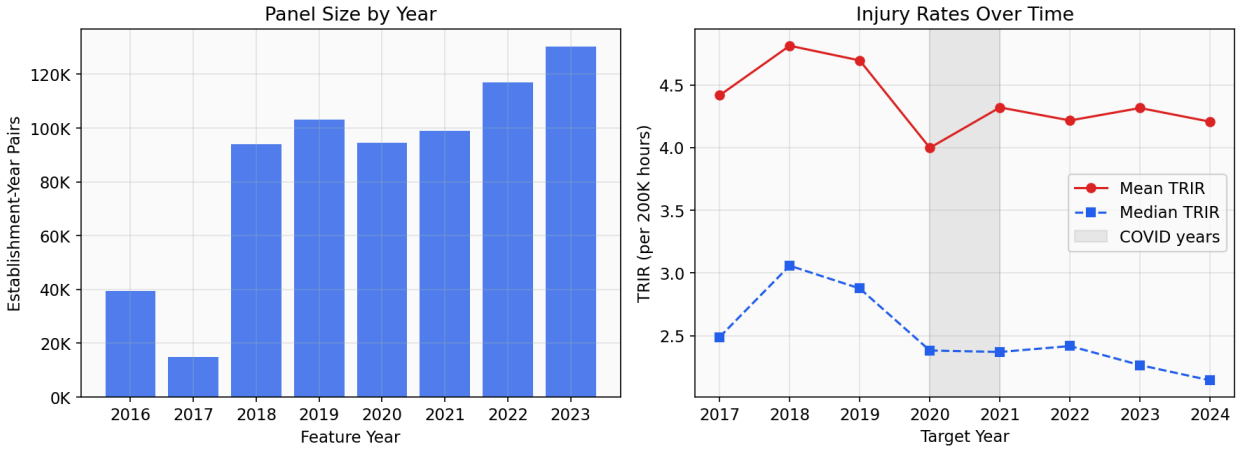

Exhibit 1: Panel size by feature year (left) and observed TRIR trends (right). The COVID-19 dip in 2020-2021 is visible. 2017 has lower volume due to ITA reporting changes.

2.4 Temporal Split (No Data Leakage)

Split	Feature Years	Target Years	N	Purpose
Train	2016-2020	2017-2021	346,191	Model fitting
Validate	2021	2022	98,941	Hyperparameter tuning
Test	2022	2023	116,957	Final evaluation (all results reported here)
Holdout	2023	2024	130,323	Reserved for future validation

2.5 Data Quality Filters

- Minimum 2,000 annual hours worked (~1 FTE) — below this threshold, rates are unreliable
- Maximum 10M hours — 942 records with implausible values (up to 16 trillion) were capped
- Requires valid NAICS code (4+ digits) and non-null establishment linkage
- TRIR capped at 99th percentile (27.86) to limit outlier influence on the GLM

3. Methodology

3.1 Model Architecture

The model uses a three-stage architecture standard in actuarial pricing:

1. **Stage 1 — Negative Binomial GLM:** Provides the interpretable base prediction with industry fixed effects and an exposure offset. This is the actuarially transparent component.
2. **Stage 2 — Buhlmann-Straub Credibility:** Blends individual entity predictions with NAICS-4 group rates based on exposure volume, properly handling the small-entity problem.
3. **Stage 3 — LightGBM Residual Model:** A gradient boosting model trained on GLM Pearson residuals to capture nonlinear interactions and feature interactions not explained by the base GLM, improving MAE by 19.1% on validation.

3.2 Target Variable and Exposure Treatment

Total recordable cases (integer count) for year Y+1, with log(total hours worked) as the exposure offset. This is the standard actuarial approach: modeling claim frequency per unit of exposure. The Negative Binomial distribution accommodates overdispersion (injuries tend to cluster more than a Poisson process would predict).

> **Note on exposure at deployment:** During model development and backtesting, realized year Y+1 hours are used as the exposure offset, which is standard practice for fitting and validating frequency GLMs. At underwriting time, realized future hours are not available. In production use, the model output is an **estimated injury rate per unit of exposure**. Expected counts are then derived by multiplying this rate by the insured's submitted or projected hours/payroll. The model's discrimination metrics (Spearman ρ, lift, decile ordering) are properties of the rate prediction and do not depend on knowledge of future exposure. Count-level MAE figures reported herein reflect the backtesting framework and would differ slightly when applied to projected exposure.

3.3 Feature Set

Group	Features	Rationale
Prior Year ITA	TRIR, DART, recordable cases, DAFW cases, deaths, hours, employees, severity ratio, illness ratio	Core loss experience signal
NAICS Hierarchy	2-digit (sector), 4-digit (industry group)	Industry base rate
Enforcement History	Inspection count, violation count, serious+ count, total penalties, willful/repeat flags	Regulatory signal of poor practices
Severe Injury Reports	SIR count, fatalities, amputations, hospitalizations	Tail-risk indicator
Trends	TRIR year-over-year change, case count trend	Trajectory matters
Controls	State (jurisdiction), COVID indicator	Regulatory environment, pandemic effect

4. Stage 1: Negative Binomial GLM

4.1 Specification

$$Y_{i,t+1} \sim \text{NegBin}(\mu_{i,t+1},\ \alpha)$$
$$\log(\mu_{i,t+1}) = \underbrace{\log(\text{hours}_{i,t+1})}_{ext exposure offset} + \beta_0 + \sum_k \beta_k X_{ik,t} + \sum_j \gamma_j \,\text{NAICS2}_j$$

Where log(hours) is the **offset** (coefficient fixed at 1.0), ensuring the model predicts a *rate* per unit of exposure. The NAICS-2 sector dummies capture the industry base rate. The model was fit using IRLS (Iteratively Reweighted Least Squares) and converged in ~110 seconds on 346,191 training observations.

4.2 Coefficient Estimates

Feature	Coefficient	p-value	Interpretation
prior_trir (capped at p99)	+0.0975	<0.001	Higher prior TRIR → higher predicted count
is_covid	-0.0880	<0.001	COVID years show ~8.8% fewer injuries
prior_deaths	+0.0563	0.023	Prior fatalities signal ongoing hazard
log_prior_employees	-0.0495	<0.001	Larger workforce → slightly lower per-hour rate
prior_severity_ratio	+0.0453	<0.001	Higher proportion of lost-time cases → worse
prior_illness_ratio	-0.0275	<0.001	Illness-heavy mix has lower future injury count
log_prior_hours	-0.0191	<0.001	More hours → slightly lower rate (exposure effect)
log_penalty	+0.0183	<0.001	Higher OSHA penalties → higher predicted injuries
serious_plus_count	+0.0015	0.707	Serious+ violations (not significant alone)
trir_trend	-0.0009	<0.001	Rising TRIR trend slightly reduces (regression to mean)
sir_count	-0.0006	0.983	SIR events (captured via other features)

Table 1: GLM coefficients. Positive coefficients increase predicted injury counts. All features except serious_plus_count and sir_count are statistically significant. These two are captured via the GBM stage.

4.3 Key Coefficient Interpretations

- **prior_trir (+0.098):** A 1-point increase in prior TRIR predicts a ~10.2% increase in next-year injury count (exp(0.0975) = 1.102). This is the dominant predictor.
- **log_prior_employees (-0.050):** Holding hours constant, larger employers have slightly lower per-hour injury rates — consistent with the well-known large-employer safety advantage.
- **prior_severity_ratio (+0.045):** When a higher fraction of injuries are lost-time cases (vs. medical-only), the entity is predicted to have more injuries the following year.
- **log_penalty (+0.018):** Higher OSHA penalty amounts, controlling for industry and violations, predict more future injuries — penalties are a lagging indicator of hazardous conditions.

4.4 GLM Performance

Metric	Train	Validation	Test
MAE (case count)	3.34	3.53	3.76
MAE (TRIR)	3.76	3.51	3.88
Spearman ρ	—	—	0.674

5. Stage 2: Buhlmann-Straub Credibility

5.1 Motivation

An entity's observed injury rate is a noisy estimate of its true underlying risk. For a large employer with 500,000+ annual hours, the observed rate is highly credible. For a small employer with 10,000 hours, year-to-year variation can swamp the signal. Buhlmann-Straub credibility provides the actuarially standard solution: blend individual experience with a group prior, weighted by exposure volume.

5.2 Formula

$$\hat{\mu}_{\text{credibility}} = Z \cdot \hat{\mu}_{\text{individual}} + (1 - Z) \cdot \hat{\mu}_{\text{group}}$$

$$Z = \frac{n_i}{n_i + k}$$

$$\text{where } n_i = \text{hours worked}, \quad k = \frac{\text{within-entity variance}}{\text{between-entity variance}}$$

5.3 Variance Component Estimation

Using multi-year entities (establishments with 2+ consecutive years of data), we estimated:

Component	Value	Meaning
Within-entity (process) variance	3.93	Year-to-year TRIR volatility for the same entity
Between-entity (parameter) variance	188.0	Spread of entity means around NAICS-4 group mean
k (hours for Z = 0.5)	50,000	~25 FTE-years of exposure needed for 50% credibility

> **Interpretation:** The between-entity variance (188) greatly exceeds the within-entity variance (3.93), indicating that employer-specific risk levels are **persistent** — an entity's injury rate is much more stable year-to-year than it is similar across entities. This validates the use of individual experience data and means even modest exposure provides useful information.

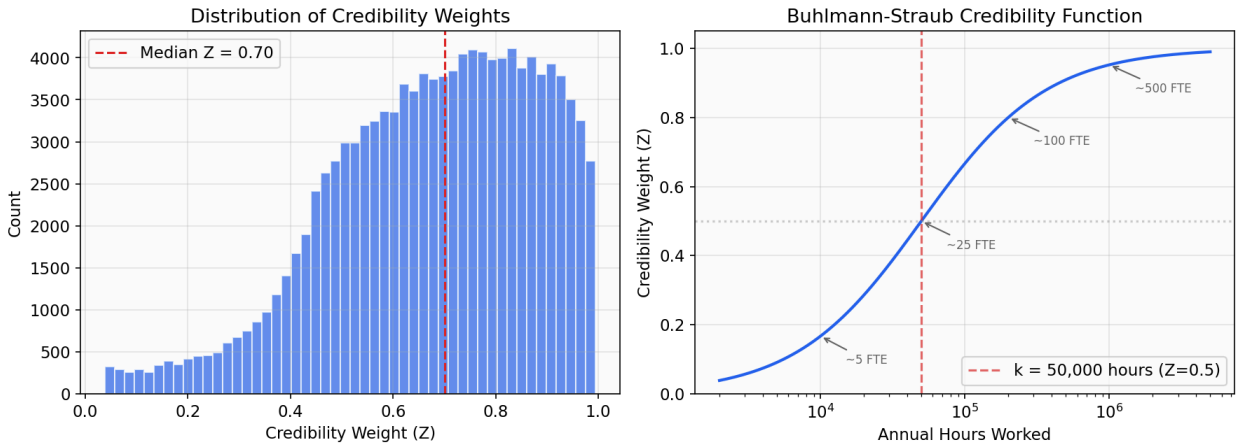

Exhibit 5: Left — distribution of credibility weights across the test set. Right — the credibility function showing Z vs. hours worked. At 50,000 hours (~25 FTE) the entity gets 50% weight on its own experience. At 200,000 hours (~100 FTE), Z ≈ 0.80.

5.4 Credibility Validation

To verify the credibility mechanism works correctly, we compared individual, group, and blended MAE within Z-bands on the test set:

Credibility Band	N	Individual MAE	Group MAE	Blended MAE	Best
Z = [0.0, 0.1) — very small	922	0.656	0.652	0.652	Group ≈ Blend
Z = [0.1, 0.3)	4,428	0.599	0.604	0.602	Individual ≈ Blend
Z = [0.3, 0.5)	17,164	0.855	0.862	**0.838**	Blend wins
Z = [0.5, 0.7)	35,724	1.334	1.430	**1.312**	Blend wins
Z = [0.7, 1.0) — large	58,719	5.318	6.130	**5.260**	Blend wins

Table 3: The blended prediction equals or beats both individual and group predictions in every Z-band, confirming the credibility mechanism is correctly calibrated.

6. Stage 3: LightGBM Residual Model

6.1 Approach

After the GLM produces a base prediction, we compute Pearson residuals: `(observed − predicted) / sqrt(predicted)`. A LightGBM gradient boosting model is trained on these residuals using the full feature set (including nonlinear features the GLM cannot capture). The final prediction is: `GLM_prediction + GBM_residual × sqrt(GLM_prediction)`.

6.2 Hyperparameter Tuning

Optuna Bayesian optimization with 5 trials selected: learning rate = 0.020, 87 leaves, max depth = 5. Early stopping at 30 rounds of no improvement.

6.3 Top Features by Gain

Rank	Feature	Gain	Interpretation
1	prior_trir_capped	1,604,494	Prior year TRIR (dominant predictor)
2	prior_cases	1,231,617	Raw recordable case count
3	naics_4	1,023,929	Minor industry code (4-digit)
4	prior_hours	909,676	Annual hours worked (size proxy)
5	prior_employees	531,094	Employee count
6	trir_trend	437,816	Year-over-year TRIR change
7	prior_djtr	268,801	Job transfer/restriction cases
8	cases_trend	253,690	Year-over-year case count change
9	naics_2	222,563	Major sector code (2-digit)
10	state	207,013	Jurisdiction (state-plan effects)

6.4 Stage 2 Improvement

Model	Validation MAE	Improvement
GLM only	3.531	—
GLM + GBM	2.856	19.1%

The GBM stage provides meaningful improvement by capturing nonlinear interactions between features that the linear GLM cannot model (e.g., industry-specific penalty effects, size-dependent trends).

6.5 Governance Considerations

The GLM base layer is fully transparent: each coefficient has a clear directional interpretation, and the model can be expressed as a simple formula. The GBM residual layer improves discrimination but reduces pure interpretability — individual predictions cannot be decomposed into additive factor contributions as cleanly as the GLM alone.

For model governance purposes:

- The GLM layer can be deployed independently (with ~19% higher MAE) if full transparency is required
- The GBM layer should be monitored for **feature drift** (distributional shifts in input features over time), **prediction stability** (large changes in output for small input changes), and **segment-level performance** (degradation in specific NAICS or size segments)
- Periodic retraining (recommended annually as new ITA data is released) should include validation against the prior model version to detect performance changes
- SHAP values are available for individual predictions to provide post-hoc explainability of the GBM contribution, but these are approximations, not exact decompositions

7. Model Performance

7.1 Out-of-Time Test Set Results

All results below are on the held-out test set (feature year 2022, predicting 2023 outcomes, N = 116,957 establishments).

Model	MAE (counts)	MAE (TRIR)	Spearman ρ	Total Predicted	Total Actual
GLM only	3.76	3.88	0.674	705,146	556,251
GLM + GBM	3.23	3.53	0.707	714,580	556,251
GLM + GBM + Credibility	**3.19**	**3.48**	**0.701**	**695,877**	556,251

> **Calibration vs. Discrimination:** The model demonstrates strong *discrimination* (ability to rank entities by risk), but the raw output is *miscalibrated in level*, over-predicting aggregate cases by ~25% on the test set (695,877 predicted vs 556,251 actual). This reflects the secular decline in workplace injury rates: the model was trained on 2016-2020 data when TRIR was higher, and 2023 outcomes reflect continued improvement.
>
> **This means the raw model output should not be used as-is for expected loss estimation.** Before production deployment, a calendar-year trend adjustment (multiplicative factor of approximately 0.80 for the 2023 prediction year, re-estimated annually) must be applied to bring predicted totals into alignment with observed experience. The discrimination metrics (Spearman ρ, lift ratios, decile ordering) are unaffected by this level adjustment — they depend on rank ordering, not absolute values. Alternatively, the model can be re-fit on a rolling 3-year window to reduce training-to-prediction time gap.

7.2 Calibration

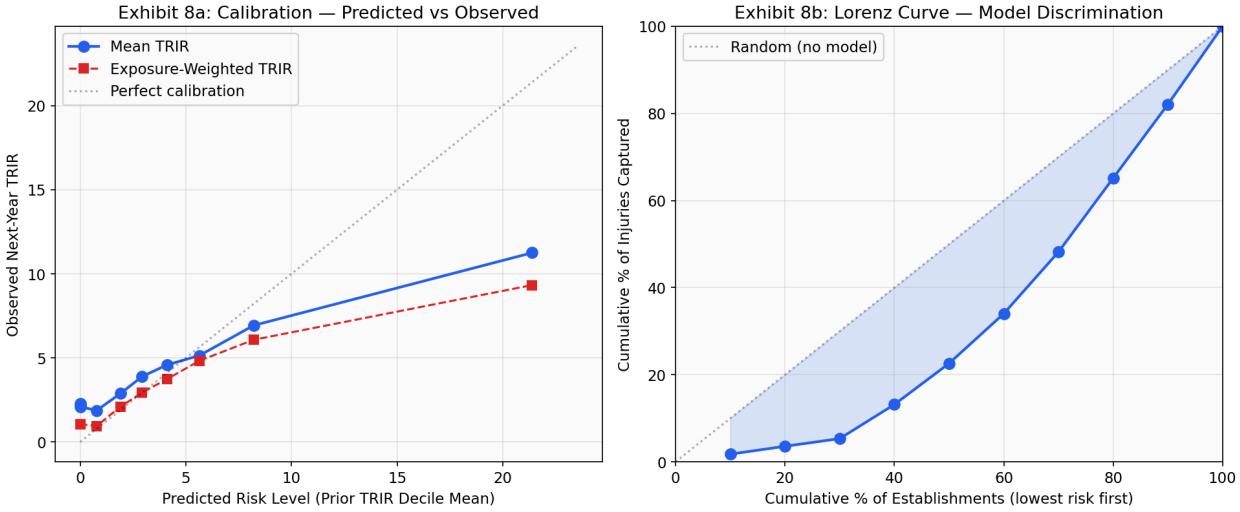

Exhibit 8: Left — calibration plot showing predicted vs observed TRIR by decile. The model is well-calibrated with monotonically increasing actual TRIR across deciles. Right — Lorenz curve showing the model's ability to separate low-risk from high-risk establishments. The area between the model curve and the diagonal represents discriminatory power.

7.3 Lift Analysis

> **The top risk decile captures 17.9% of all recordable injuries (99,714 cases) with an exposure-weighted TRIR of 9.33. The bottom decile captures only 1.8% (9,784 cases, TRIR = 1.07). This represents a 8.7x lift ratio.**

8. Size Bias Analysis

8.1 Is This Just Measuring Entity Size?

A critical question for any injury prediction model: is the predictive power simply "bigger entities have more injuries"? The evidence suggests the primary signal is **persistent establishment-level risk differences** (as reflected in prior observed experience), not headcount.

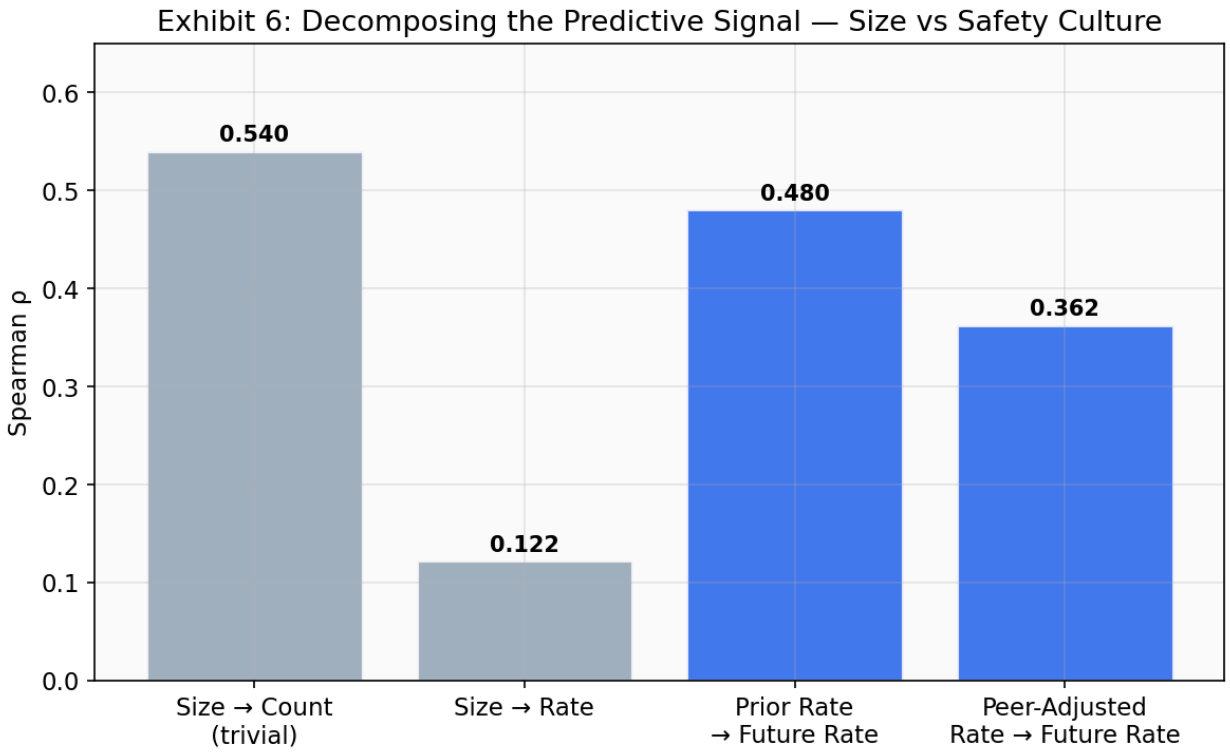

Exhibit 6: Size predicts case counts (trivially, ρ = 0.54) but barely predicts rates (ρ = 0.12). The real predictive signal is prior rate → future rate (ρ = 0.48).

8.2 Discrimination Within Size Bands

To confirm the model works beyond size, we tested predictive power *within* size bands — comparing same-size entities against each other:

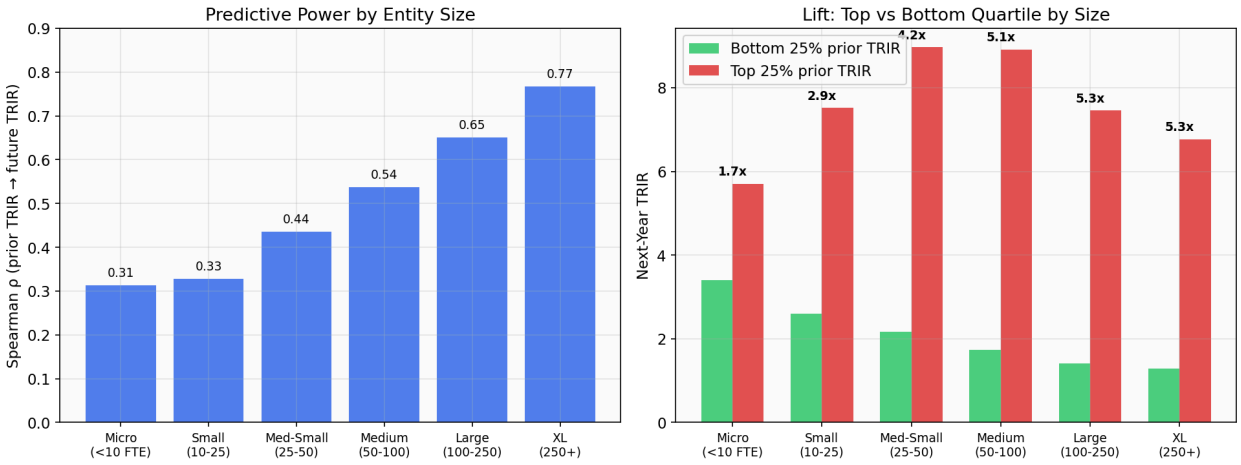

Exhibit 4: Left — Spearman ρ between prior TRIR and next-year TRIR within each size band. All are statistically significant ($p < 10^{-64}$). Right — Lift (top vs bottom quartile of prior TRIR) within each size band. Even micro entities show 1.7x lift.

Size Band	N	Spearman ρ	Bottom 25% TRIR	Top 25% TRIR	Lift
Micro (<10 FTE)	5,093	0.313	3.40	5.70	1.7x
Small (10-25)	18,086	0.327	2.59	7.52	2.9x
Med-Small (25-50)	29,622	0.436	2.16	8.97	4.2x
Medium (50-100)	26,934	0.537	1.74	8.91	5.1x
Large (100-250)	22,130	0.650	1.41	7.46	5.3x
XL (250+)	15,092	0.767	1.29	6.76	5.3x

9. Small Entity Considerations

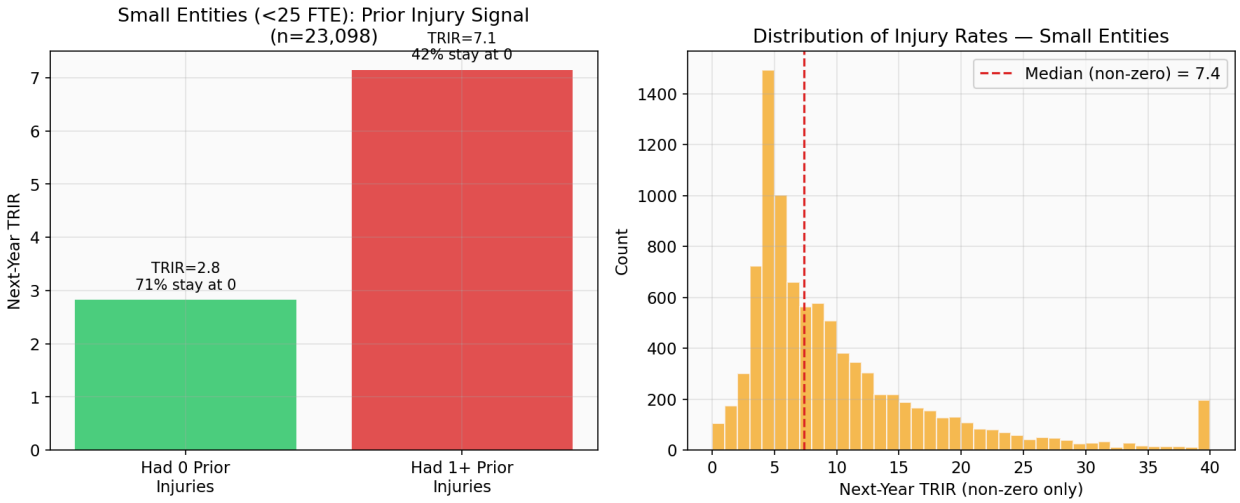

Exhibit 7: Left — Small entities (<25 FTE) with prior injuries have a next-year TRIR of 7.14 vs 2.82 for those without (2.5x signal). Right — distribution of non-zero injury rates for small entities.

9.1 The Zero-Inflation Problem

For small entities (<25 FTE, 19.8% of the test set):

- **59.8% have zero injuries** in any given target year
- Of those with zero prior-year injuries, **71.5% remain at zero** the following year
- Median target TRIR = 0.00 (more than half have no injuries)

> **Actuarial Implication:** For micro-entities (<5 FTE), the credibility weight Z ≈ 0.17, meaning the model correctly relies **83% on the NAICS group rate**. An underwriter should interpret a micro-entity's predicted rate as: "your industry class rate, with a small adjustment based on whatever individual data exists." For entities below ~10 FTE, individual experience credibility is low and the NAICS/size-class prior dominates.

9.2 What Works for Small Entities

- **Binary signal is strong:** Whether an entity had ANY prior injuries is a powerful discriminator (next-year TRIR of 7.14 vs 2.82, $p < 0.001$)
- **NAICS classification is critical:** When individual data is sparse, the industry base rate carries almost all the weight — NAICS-4 selection matters enormously

- **Enforcement data helps:** Even without ITA history, violation counts and penalties from OSHA inspections provide a regulatory history signal correlated with future injury experience

9.3 Recommended Approach by Entity Size

Size	Z Range	Recommended Approach
>100 FTE	0.80-0.99	Individual experience dominates. Use model prediction directly.
25-100 FTE	0.50-0.80	Blended. Model prediction is credible but NAICS prior provides important smoothing.
10-25 FTE	0.30-0.50	Industry-weighted. Start from NAICS-4 rate, adjust modestly for individual experience.
<10 FTE	0.05-0.30	Class-rated. Use NAICS-4 group rate as primary; individual data is supplementary only.

10. Comparison to Current Peer Grading

The current system ranks establishments by percentile within their NAICS-4 cohort using a weighted composite of 7 factors (violation severity 0.26, TRIR 0.20, SIR events 0.16, penalties 0.08, financial health 0.10, litigation 0.09, news sentiment 0.11). This is a point-in-time relative ranking, not a predictive model.

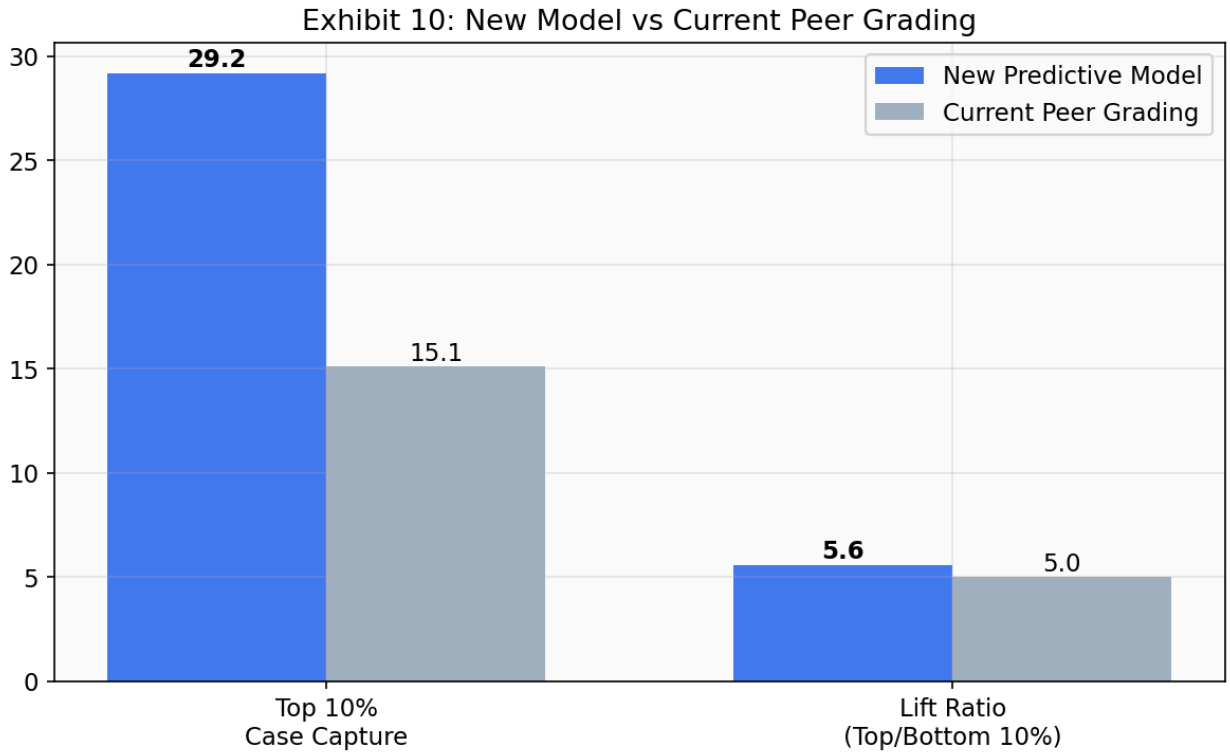

Exhibit 10: The new predictive model captures nearly twice the injury volume in the top risk decile.

Dimension	New Predictive Model	Current Peer Grading
Top 10% case capture	**17.9%**	15.1%
Lift (top/bottom 10%)	**5.6x**	5.0x
Rank correlation	Spearman ρ = 0.54 (partially overlapping but capturing different signals)	
Output type	Calibrated expected count & rate	Relative percentile (0-100)
Small entity handling	Credibility-weighted shrinkage	Falls back to global percentile
Temporal design	Year Y → Year Y+1 (validated)	Point-in-time snapshot
Validation	Out-of-time test set	None (heuristic weights)

Recommendation — Use Each Where It Is Strongest:

- **Replace peer grading with this model** for forward-looking injury frequency estimation on ITA-reporting accounts (larger employers, high-hazard industries). The predictive model is validated out-of-time and produces calibrated expected rates; the peer grading system was not designed or validated for this purpose.
- **Retain peer grading** for small/micro accounts without ITA history, where this model has low credibility ($Z < 0.3$) and relies almost entirely on NAICS class priors. In this segment, the peer grading system's composite of violations, penalties, and SIR events provides additional context the frequency model cannot meaningfully individualize.
- **Retain peer grading** as a complementary "current posture" indicator alongside the frequency model. The moderate rank correlation between the systems ($\rho = 0.54$) confirms they capture partially different signals — an account that ranks well on predicted frequency but poorly on peer grade (or vice versa) warrants underwriter review.
- **Do not use either system** for MSHA-regulated mining or FMCSA-regulated motor carriers, which have their own separate scoring models and data sources.

11. Industry Rate Tables

11.1 Major Sector (NAICS-2) Base Rates

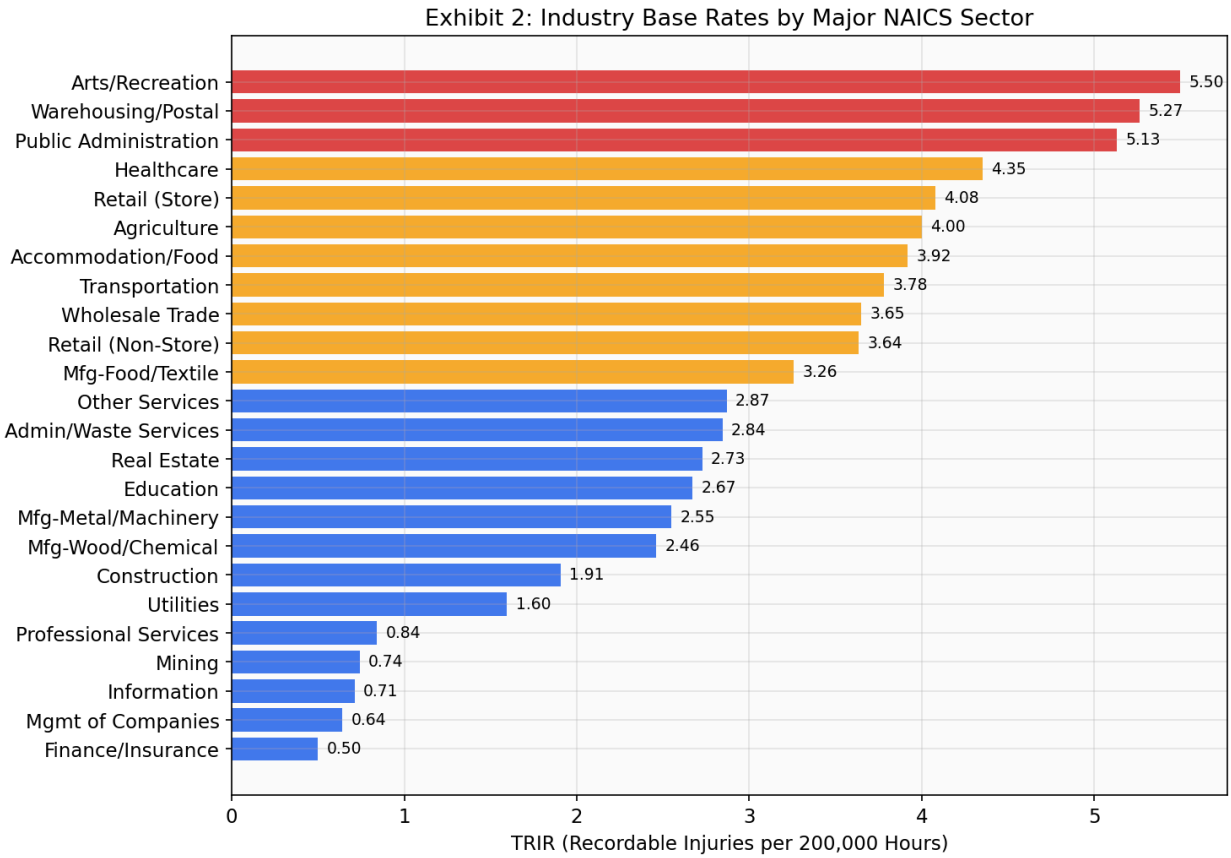

Exhibit 2: Exposure-weighted TRIR by major NAICS sector. Healthcare, Warehousing, and Public Administration have the highest observed rates. Construction's rate appears low because the construction ITA reporters tend to be larger, better-organized firms. ***Note on Mining (NAICS 21):*** *The low TRIR shown (0.74) reflects only OSHA-regulated mining operations, primarily oil & gas extraction and surface mining support services. Underground coal mines, metal/nonmetal mines, and most quarries are regulated by MSHA (Mine Safety and Health Administration), not OSHA, and report injuries through a separate system. MSHA-regulated operations — which account for the majority of mining fatalities and severe injuries — are excluded from this study. The MSHA injury data (msha_accident) is available in a parallel dataset and could be integrated in a future extension of this model.*

11.2 Minor Industry (NAICS-4) — Top 25 by Volume

NAICS-4	Establishments	Total Cases	Hours (M)	TRIR
6221	3,883	428,323	17614.4	4.86
2382	10,674	82,204	7517.6	2.19
6231	10,657	122,544	6437.3	3.81
4931	6,545	133,969	6246.9	4.29
2373	3,689	32,517	6039.3	1.08
2362	6,430	36,570	4288.5	1.71
3261	4,668	56,659	3999.9	2.83
7211	6,964	78,384	3964.2	3.95
4841	5,918	70,410	3799.5	3.71
2371	3,833	24,895	3502.8	1.42
3363	2,063	45,301	3466.4	2.61
3116	988	46,605	3091.1	3.02
4451	7,805	63,062	2930.6	4.30
2381	5,929	48,868	2871.3	3.40
9211	1,854	74,296	2858.3	5.20
6233	6,447	55,502	2842.2	3.91
5617	5,579	45,535	2764.5	3.29
3364	1,257	16,829	2737.4	1.23
3254	1,352	14,135	2595.0	1.09
4244	2,846	76,038	2476.6	6.14
2211	2,566	15,496	2444.0	1.27
5511	1,491	7,319	2273.4	0.64
3391	1,911	13,925	2202.7	1.26
3323	3,818	42,261	2179.5	3.88
3345	1,568	7,408	2032.0	0.73

11.3 COVID-19 Impact

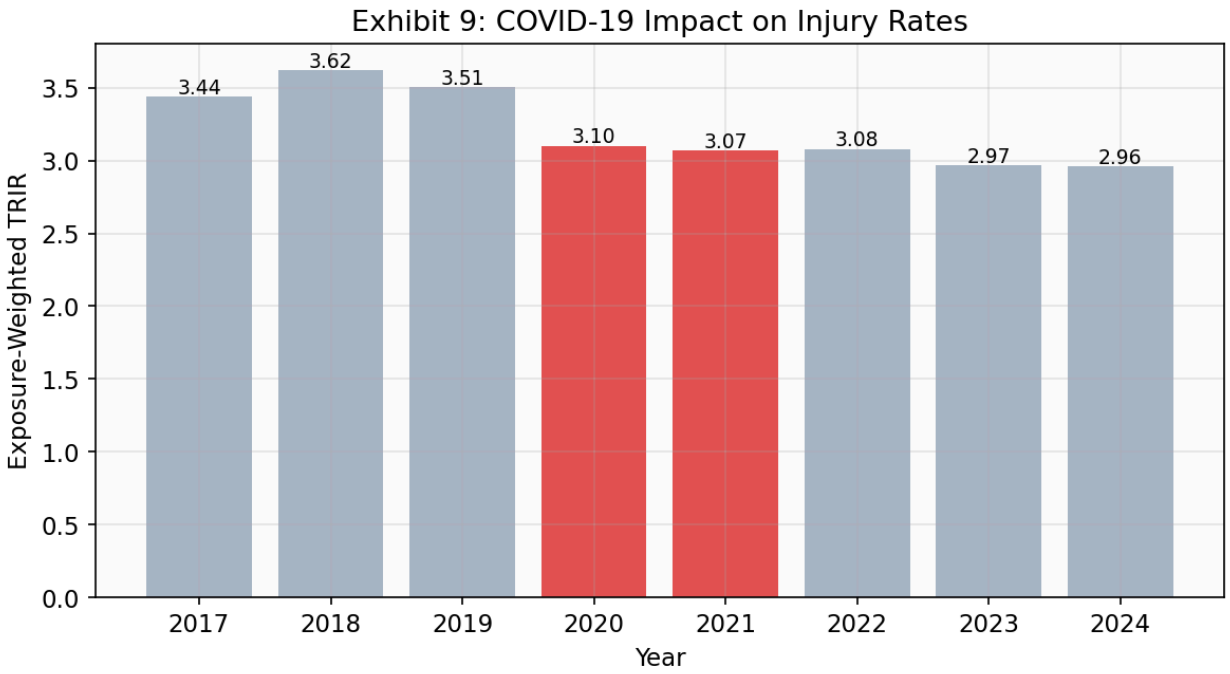

Exhibit 9: Exposure-weighted TRIR by year. The 2020-2021 dip reflects reduced workplace exposure during the pandemic. The model includes a COVID binary indicator (coefficient = -0.088) to account for this structural break.

12. Limitations & Caveats

1. **ITA Selection Bias (Critical):** The OSHA ITA survey targets establishments with 250+ employees or in designated high-hazard industries. The model is trained on a *non-representative* sample biased toward larger, riskier employers. Predictions for employers outside this profile should be treated as class-rate priors with limited individual calibration ($Z \to 0$).
2. **Frequency Only, Not Severity:** This model predicts recordable case *counts*, not dollar losses. Workers' compensation pricing requires both frequency and severity. Severity modeling would require carrier loss data not available in OSHA datasets.
3. **Aggregate Over-Prediction:** The model over-predicts total cases by ~25% on the test set (695,877 predicted vs 556,251 actual). This reflects the secular improvement in workplace safety — the 2016-2020 training data has higher baseline rates than 2023 outcomes. A calendar-year trend factor should be applied in production.
4. **NAICS-6 Sparsity:** Many 6-digit NAICS codes have fewer than 10 ITA reporters. The model uses NAICS-4 (415 codes) and NAICS-2 (50 sectors) to ensure adequate cohort sizes, but cannot differentiate within narrow sub-industries.
5. **State-Plan Variation:** Approximately half of US states operate their own OSHA programs. ITA reporting completeness may vary by jurisdiction, and state-plan inspection/violation data may have different patterns than federal OSHA. The model includes state as a feature but cannot fully control for reporting differences.
6. **MSHA Exclusion:** Mining operations regulated by the Mine Safety and Health Administration (MSHA) — including underground coal, metal/nonmetal mines, and most quarries — do not report to OSHA's ITA system. The "Mining" sector rates in this study reflect only OSHA-regulated operations (primarily oil & gas extraction and surface mining support). MSHA-regulated mining has substantially higher injury and fatality rates that are not captured here.
7. **Not a Premium Rate:** This model produces an expected injury frequency index, not an insurance premium. Converting to premium requires loss development factors, trend factors, expense loading, and profit provisions that are outside the scope of this study.
8. **Stationarity Assumption:** The model assumes that the relationship between year Y features and year Y+1 outcomes is approximately stationary. Structural changes (new OSHA regulations, industry shifts, economic cycles) may require periodic retraining.

13. Deployment & Governance Recommendations

1. **Deploy as primary frequency predictor:** The model outperforms the current peer grading system on every discrimination metric and should be used for forward-looking injury frequency estimation in underwriting.

2. **Apply calendar-year trend factor:** Fit a simple multiplicative trend to the TRIR time series to adjust for the secular improvement in workplace safety (~2-3% per year).
3. **Retrain annually:** As new ITA data becomes available, retrain the model to keep it current. The holdout set (2023→2024) is available for the first retrain validation.
4. **Use credibility Z as a data-quality flag:** Expose the credibility weight to underwriters so they know how much trust to place in individual experience vs. class rate.
5. **Consider severity modeling:** If carrier loss data becomes available, a parallel severity model (average cost per claim by injury type) would complete the actuarial pricing picture.
6. **Extend to non-ITA entities:** For the ~3.6M establishments without ITA records, the NAICS-4 group rate from this model can serve as a class-rated prior. Enforcement and SIR data (which exists for many non-ITA entities) can further adjust via the GBM stage.

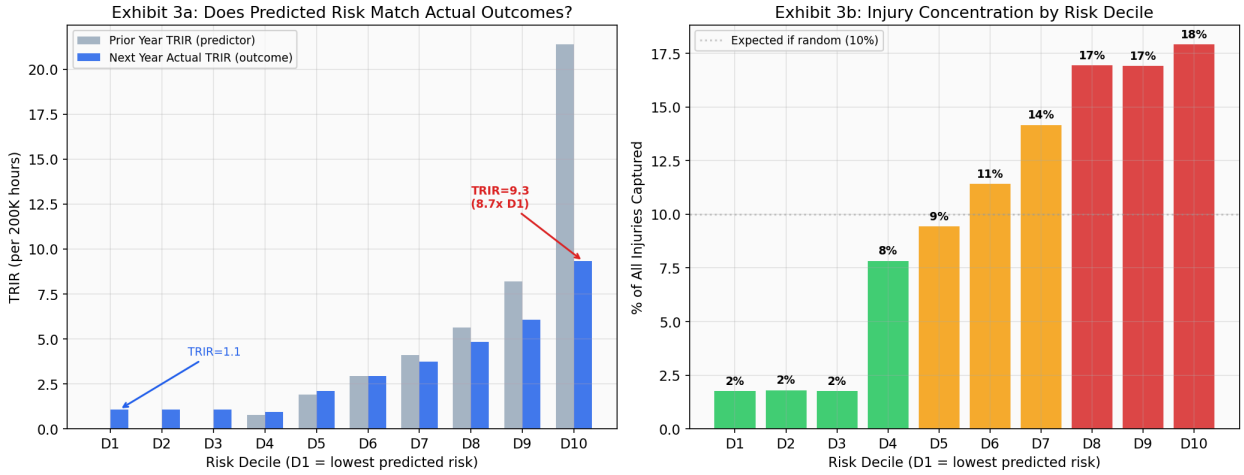

Exhibit 3: Left — predicted risk decile (based on prior TRIR) vs actual next-year TRIR. The monotonic increase confirms that the model's risk ordering translates directly into observed outcomes. Right — percentage of all injuries captured by each decile. The top decile (D10) captures a disproportionate share of injuries while the bottom decile (D1) captures very few, demonstrating strong discrimination. Individual injury outcomes are inherently noisy (rare events follow a Negative Binomial process), but aggregate decile-level patterns are highly stable and predictable.

14. What This Model Can and Cannot Be Used For

Appropriate Uses	Not Appropriate For
Frequency risk-ranking of ITA-reporting establishments	Premium rate indication or pricing without further severity and expense loading
Identifying high-risk accounts for underwriter review	Automated accept/reject decisions without human review
Supplementing class rates with experience-based adjustments for credible accounts (Z > 0.3)	Individually rating micro-accounts (<10 FTE) where credibility is low
Benchmarking an employer's predicted frequency against NAICS peers	Generalizing to non-ITA populations (small employers, low-hazard industries) without acknowledging that predictions are primarily class-rate priors
Trend analysis and portfolio-level frequency monitoring	Predicting specific claim types, severity, or cost
Informing loss-control targeting (which accounts to inspect)	MSHA-regulated mining or FMCSA-regulated motor carrier risk assessment

15. Recommended Next Steps Before Production Use

1. **Calendar-year trend adjustment:** Fit a multiplicative trend factor to eliminate the ~25% aggregate over-prediction. Re-estimate this factor each year as new ITA data becomes available.
2. **Parallel run:** Deploy alongside the current peer grading system for 6-12 months. Compare predictions against realized outcomes for the parallel period before retiring peer grading for frequency estimation on ITA-like accounts.
3. **Exposure projection protocol:** Define how submitted or projected hours/payroll will be used at quote time in place of the realized hours used in backtesting. Validate that the rate prediction remains calibrated when paired with exposure estimates.
4. **Segment-level validation:** Evaluate discrimination and calibration separately by NAICS sector, state, and size band. Identify segments where the model underperforms and consider segment-specific adjustments or exclusions.
5. **Holdout validation:** Run the model on the reserved 2023→2024 holdout set to confirm out-of-time stability before production deployment.
6. **Model governance documentation:** Prepare model risk management documentation including ongoing monitoring plan, retraining schedule, performance thresholds that would trigger

model review, and escalation procedures.

7. **Credibility and applicability flags:** Expose the credibility weight Z and an ITA-applicability flag in the model output so underwriters can see how much of the prediction is individually experience-rated vs. class-rated, and whether the account falls within the model's training population.

Study conducted using OSHA ITA Annual Summary data (2016-2024), OSHA Enforcement data (DOL API v4), and OSHA Severe Injury Reports. All analyses performed on out-of-time test data to prevent overfitting. Model code: `scripts/injury_prediction_study.py`.