femington

Int. J. Clin. Res. Med. Sci. (2026) 1(3):1–19
Article ID: IJCRMS-2026-170

Exploring the Relationship Between Establishment Size and Workplace Injuries: An Analysis of OSHA Data

Dr. Nandini Singh
nandinisinghbhati23@gmail.com
Independent Researcher

Abstract

This study examines the relationship between establishment size and workplace injuries using data from the Occupational Safety and Health Administration (OSHA). The analysis focuses on over 350,000 establishments, utilizing various injury metrics, including total injuries, skin disorders, respiratory conditions, and poisoning rates. Despite initial hypotheses suggesting larger establishments would exhibit lower injury rates per worker, findings reveal minimal correlation between injury rates and establishment size, evidenced by a weak negative correlation coefficient of -0.0136. Data cleaning and variable generation processes were essential to address anomalies and enhance the integrity of the analysis. Outliers significantly influenced results, highlighting the need for improved data collection and reporting standards. The study concludes that further research, potentially incorporating data from different years and sources, is necessary to understand workplace safety dynamics better, particularly regarding respiratory conditions and the impact of COVID-19. This investigation lays a foundation for future inquiries into workplace accidents and underscores the importance of addressing reporting inconsistencies in occupational safety data.

Keywords

• Workplace Safety • Establishment Size • Injury Rates • Occupational Safety and Health Administration (OSHA) • Occupational Health • Data Analysis • COVID-19 Impact

1. Introduction

1.1 The Importance of Workplace Safety

Workplace safety stands as a fundamental concern for organizations across all industries and sectors. Over the past five decades, industry standards have increasingly emphasized the reduction of serious workplace injuries and the promotion of overall occupational safety. This shift reflects growing recognition that safe working conditions are not merely a regulatory requirement but a cornerstone of organizational success and employee well-being [1].

The importance of maintaining safe work environments extends beyond the obvious moral imperative to protect workers from harm. Organizations with strong safety records benefit from reduced workers' compensation costs, lower insurance premiums, decreased absenteeism, and enhanced productivity. Furthermore, a reputation for workplace safety can serve as a competitive advantage in attracting and retaining talent, as workers increasingly prioritize their health and safety when selecting employers [2].

Despite these clear benefits, workplace injuries remain a persistent challenge across many industries. Understanding the factors that contribute to injury rates is essential for developing effective prevention strategies and allocating resources efficiently. Recent work on AI-driven real-time monitoring systems

© 2026 Femington Academic Publishing. All rights reserved.

illustrates the growing interest in technological solutions to workplace safety challenges [3]. This study focuses on one potentially important factor: the size of the establishment where work occurs.

1.2 Establishment Size as a Potential Determinant

Establishment size may influence workplace safety through multiple mechanisms. Larger establishments might benefit from economies of scale in implementing safety programs, dedicating specialized personnel to safety management, and investing in safety equipment and training. They may also face greater scrutiny from regulators and have more formalized safety protocols [4].

Conversely, smaller establishments might lack the resources to implement comprehensive safety programs or to hire dedicated safety professionals. However, they may benefit from simpler operations, closer supervisor-worker relationships, and more flexible adaptation to safety concerns. The net effect of establishment size on injury rates remains theoretically ambiguous and empirically under-examined. Research by Weil examined these dynamics within the construction industry, finding that regulatory effects and safety outcomes vary considerably across establishment types [5].

This ambiguity motivates our research question: Does establishment size correlate with workplace injury rates, and if so, in what direction and with what magnitude? Answering this question has important implications for how safety resources should be allocated across establishments of different sizes and how regulators might tailor their approaches.

1.3 The OSHA Data Resource

The Occupational Safety and Health Administration (OSHA) collects comprehensive data on workplace injuries and illnesses from private sector employers across construction, general industry, and maritime sectors. This data, gathered through employer recordkeeping requirements, OSHA inspections, and electronic reporting systems, represents one of the most extensive sources of occupational safety information available [6].

OSHA data includes information on establishment characteristics, injury types, severity measures, and employment levels. The agency's Injury Tracking Application (ITA) collects establishment-specific injury and illness data annually, providing researchers with a rich resource for analyzing workplace safety patterns. For this study, we utilize the OSHA Injury Tracking Application (ITA) dataset for calendar year 2022, which was the most recent complete dataset available at the time of analysis. This cross-sectional data includes over 350,000 establishments and more than 20 variables describing establishment characteristics, employment levels, and injury outcomes. The data set encompasses establishments of widely varying sizes, from very small firms to those with hundreds of thousands of employees. Research on workplace injury recording practices has highlighted the importance of understanding data limitations. [7].

1.4 Research Objectives and Contributions

This study pursues three primary objectives. First, we examine the overall relationship between establishment size and total injury rates across more than 350,000 establishments. Second, we disaggregate injury types to explore whether size effects differ for specific categories such as skin disorders, respiratory conditions, and poisonings. Third, we identify data quality issues and outliers that may affect the interpretation of results.

Our contributions include an analysis of the size-injury relationship using recent data, documentation of data cleaning procedures that can inform future research, and identification of areas where improved

data collection is needed. To our knowledge, this is among the first analyses to use the most recent comprehensive OSHA data for this purpose, although we cannot claim it is the first systematic analysis without a more exhaustive literature review. The findings also provide baseline information that may be relevant to understanding how the COVID-19 pandemic could have affected workplace injury reporting, particularly for respiratory conditions, though this study does not directly test that relationship.

2. Literature Review

2.1 Early Studies of OSHA Effectiveness

The effectiveness of workplace safety regulation has been a subject of scholarly inquiry since OSHA's establishment in 1970. Early studies by Viscusi and Smith examined whether OSHA inspections actually reduced injury rates [1, 4]. These investigations found that the increase in inspections during 1973 correlated with a 17 percent reduction in workplace injuries in subsequent months. However, by 1974, injury rates had rebounded to their original levels, raising questions about the persistence of inspection effects and the mechanisms through which they operate.

Viscusi's analysis highlighted the complexity of assessing regulatory impact, noting that firms may respond to inspection threats through temporary compliance efforts that do not translate into sustained safety improvements [1]. This finding underscores the importance of examining structural factors such as establishment size that may influence both baseline safety levels and responses to regulatory pressure.

Smith's contemporaneous work reached similar conclusions while also identifying variation in inspection effects across industries and establishment types [4]. His analysis suggested that larger establishments might be better positioned to implement lasting safety improvements due to their greater resources and more formalized management structures.

Gray and Scholz provided a comprehensive panel analysis of OSHA enforcement, finding that regulatory enforcement does produce measurable reductions in injury rates, though the magnitude of these effects depends on the specific enforcement strategies employed [8]. Their work demonstrated that inspections, particularly those resulting in penalties, can generate sustained compliance improvements, especially among establishments with prior violations.

Bartel and Thomas examined both direct and indirect effects of OSHA regulation, finding that the regulatory impact extends beyond inspected establishments to influence industry-wide safety practices [9]. This diffusion effect suggests that the total regulatory impact may be substantially larger than estimates based solely on inspected establishments would indicate.

2.2 Contemporary Evidence on Safety Determinants

More recent research has expanded understanding of workplace safety determinants. Haviland and colleagues provided a new estimate of OSHA inspection effects using manufacturing data from 1998-2005, finding that inspections reduce injury rates by approximately 9-10 percent, with effects persisting for several years following inspection [10]. This work confirmed earlier findings while providing more precise estimates using improved methodologies.

Li examined whether public availability of establishment injury data creates incentives for safety improvements [11]. The study found that transparency regarding injury rates can motivate establishments to enhance safety practices, though effects vary across industries and establishment types.

Johnson's research on the mining industry demonstrated how external economic factors can influence workplace safety [12]. When demand for resources increases suddenly, injury rates tend to rise, likely

reflecting pressures to increase production speed that may compromise safety practices. This finding highlights the importance of considering industry-specific dynamics when analyzing injury patterns.

Shore conducted a firm-level analysis of OSHA's impact on workplace injuries, finding that the effects of inspections vary substantially by firm size and industry sector [13]. Smaller firms showed greater responsiveness to inspection activity, while larger firms demonstrated more consistent baseline safety performance regardless of inspection status.

2.3 Economic Incentives for Safety

Levine and colleagues conducted a randomized study of OSHA inspections, providing rigorous evidence on the causal effects of enforcement [2]. Their research found that random inspections reduced injuries by approximately 9 percent with no detectable negative employment effects. This study provides strong evidence that regulatory enforcement can improve safety outcomes while also suggesting that firms may have been operating below optimal safety levels prior to inspection.

The economic incentives facing firms regarding workplace safety have been analyzed by several researchers. Firms that create safer work environments may be able to pay lower wages, as workers accept reduced compensation in exchange for reduced injury risk [2]. This compensating differential creates a market-based incentive for safety improvements that operates alongside regulatory requirements.

Keiner and Leeth provide a comprehensive overview of the economics of occupational and product risks [14]. Their analysis situates workplace safety within broader frameworks of risk regulation, emphasizing the trade-offs between safety investments and other competing uses of firm resources.

Boden, Ruser, and Punnett conducted a thorough review of workplace injury and illness recording practices, identifying systematic reporting biases that may affect the quality of OSHA data and other administrative records [7]. Their work highlights the need for careful data cleaning and validation when analyzing occupational injury data, as reporting practices vary significantly across establishments and industries.

Mendeloff and Gray explored the mechanisms through which OSHA inspections lead to injury reductions, finding that inspections primarily work through increasing employer awareness of hazards and motivating voluntary compliance improvements rather than through penalty imposition alone [15]. This suggests that educational and advisory components of inspections may be as important as enforcement elements.

Kniesner and Leeth provided comprehensive simulations of workplace safety policy options, demonstrating that the optimal mix of regulatory approaches depends on industry characteristics and establishment features [16]. Their work provides a framework for understanding how establishment size might interact with policy effectiveness.

2.4 Technological and Data-Driven Approaches to Safety

Beyond traditional regulatory and economic analyses, recent scholarship has explored technological and data-driven approaches to workplace safety. Dhanani examined smart safety systems that leverage AI and big data for real-time construction workforce management, demonstrating how predictive analytics can identify at-risk workers and hazardous conditions before injuries occur [17]. Such systems represent a departure from reactive inspection-based models toward continuous, data-informed safety management.

The governance of data underpinning these systems has also received attention. Morthala established a data governance framework aimed at enhancing decision efficacy in Industry 4.0 environments, emphasizing that the quality and reliability of safety data depend critically on institutional structures

for data collection, validation, and use [18]. This perspective is particularly relevant to administrative datasets such as OSHA's, where reporting inconsistencies remain a persistent concern.

Methodological advances in analytics have further expanded the toolkit available to safety researchers. Amin and Shukla optimized in-database analytics for dynamic data exploration and predictive insights, showing how large-scale organizational data can be queried and modeled more efficiently without extensive data movement [19]. Katkar enhanced fault diagnosis in telemetry datasets through advanced visualization techniques, an approach that parallels the challenge of identifying anomalous injury reports in establishment-level data [20].

Complementing these developments, Chitukoori developed an IoT-based environmental monitoring system using Raspberry Pi for real-time data analysis, illustrating how low-cost sensing infrastructure can support continuous safety surveillance in industrial settings [21]. Agarwal argued for the centrality of insightful analytic visualization in preparing data for AI projects, noting that visual pathways to data preparedness can substantially improve downstream model reliability [22]. Abraham bridged big data machine learning and human insight through visual interaction, demonstrating that human-in-the-loop interfaces can surface patterns that purely automated pipelines may overlook [23]. Finally, Syed proposed semantic synthesis of analytical pipelines, automating workflow composition for predictive data science and thereby reducing the manual effort required to construct reproducible safety analytics [24].

2.5 Gaps in the Literature

Despite extensive research on workplace safety determinants, relatively little attention has been paid to the relationship between establishment size and injury rates. Existing studies have examined size as a control variable rather than as a primary focus of analysis. This gap is surprising given the theoretical reasons to expect size effects and the practical implications for targeting safety resources.

Our literature review also revealed limited research using the most recent OSHA data, which includes the period of the COVID-19 pandemic. The pandemic may have affected workplace safety in multiple ways through direct health risks, changes in work arrangements, and disruptions to safety practices. Understanding these effects requires analysis of recent data, which this study provides.

Methodological advances in other fields, including ensemble learning and systematic evaluation frameworks, suggest opportunities for more sophisticated analysis of multi-factor safety determinants. Stacking and metalearning techniques could potentially identify complex interactions between establishment characteristics, industry factors, and safety outcomes that traditional regression approaches might miss. However, such methods are beyond the scope of the present study, which employs straightforward correlation and descriptive analysis. Future research may benefit from applying these advanced techniques to occupational safety data [25].

Recent methodological developments in anomaly detection and diagnostic applications also offer promising tools for identifying reporting inconsistencies and data quality issues in administrative datasets like OSHA's [26]. These techniques could help researchers systematically identify problematic observations and improve the reliability of analyses using OSHA data.

3. Theoretical Analysis

3.1 Economies of Scale in Safety Provision

The hypothesis that larger establishments experience fewer injuries per worker rests on the concept of economies of scale in safety provision. Safety investments often involve fixed costs that can be spread

Table 1: Workplace Safety Literature

Author	Focus	Finding
Viscusi (1979)	OSHA	Effects fade
Smith (1979)	Insp.	Vary by industry
Levine et al. (2012)	Random insp.	Fewer injuries, no loss
Li (2019)	Data	Improves safety
Li (2020)	Unions	Minimal impact
Johnson (2019)	Mining	Demand ↑ injuries

across more workers in larger establishments. A comprehensive safety training program, for example, requires similar development costs regardless of whether it serves 50 or 500 workers. The per-worker cost of such programs is substantially lower for larger establishments, potentially enabling them to implement more extensive safety measures [16].

Similarly, larger establishments may find it cost-effective to employ dedicated safety professionals whose sole responsibility is to identify hazards, develop safety protocols, and ensure compliance with regulations. Smaller establishments typically cannot justify such specialized positions, relying instead on line managers whose primary focus is production rather than safety. Research on firm-level OSHA impacts has shown that smaller firms often lack the internal capacity for systematic safety management [13].

Economies of scale may also apply to safety equipment purchases. Bulk purchasing discounts, the ability to invest in automated safety systems, and the resources to maintain and upgrade equipment over time are all more accessible to larger establishments. These advantages could translate into systematically lower injury rates.

3.2 Regulatory Scrutiny and Compliance Costs

Larger establishments may face greater regulatory scrutiny than their smaller counterparts. OSHA inspection targeting often considers establishment size, with larger facilities more likely to be inspected. This increased probability of inspection creates stronger incentives for maintaining compliance and investing in safety improvements [5].

However, the relationship between size and regulatory pressure is not monotonic. Very small establishments may operate below OSHA's radar entirely, facing minimal inspection risk. Medium-sized establishments may face moderate inspection probability, while very large establishments may be subject to continuous monitoring and regular inspections. The incentive effects of regulatory pressure should therefore vary across the size distribution. Mendeloff and Gray's investigation of inspection mechanisms suggests that the informational and advisory components of inspections may matter more than the threat of penalties for many establishments [15].

Compliance costs also exhibit scale economies. Understanding and interpreting regulations, completing required paperwork, and maintaining records all involve fixed costs that are proportionally larger for smaller establishments. This compliance burden may divert resources from substantive safety improvements, potentially offsetting other advantages of smaller scale.

3.3 Organizational Complexity and Communication

Larger establishments are typically more complex organizationally, with multiple layers of management, specialized departments, and formal communication channels. This complexity can both help and hinder safety efforts. Formal safety committees, written procedures, and dedicated safety personnel can systematize hazard identification and control. However, communication breakdowns between departments,

diffusion of responsibility for safety, and delays in information transmission can undermine safety efforts. Katkar's work on fault diagnosis in telemetry data illustrates how visualization can help organizations detect and trace failures across complex operational systems, an approach directly relevant to safety monitoring in large establishments [20].

Smaller establishments benefit from simpler organizational structures. Communication is often direct and informal, allowing rapid identification and correction of hazards. Supervisors have closer relationships with workers and may be more aware of individual safety concerns. However, this informality can also mean that safety practices depend heavily on individual discretion rather than systematic procedures.

3.4 Workforce Characteristics

Establishment size correlates with workforce characteristics that may independently affect injury rates. Larger establishments are more likely to employ unionized workers, who may have greater awareness of safety rights and more willingness to report injuries. However, as Li found, unionization does not strongly predict injury rates [11].

Larger establishments may also have more diverse workforces, including workers in different occupations with varying injury risks. Analyzing overall injury rates without accounting for occupational mix could obscure size effects. Our approach of examining specific injury types helps address this concern, as different injury types are associated with different occupational exposures.

3.5 Theoretical Summary

The theoretical relationship between establishment size and injury rates is ambiguous. Multiple mechanisms operate in different directions, and their net effect is an empirical question. Our analysis aims to provide evidence on this relationship using comprehensive OSHA data.

4. Data

4.1 OSHA Data Source and Collection

The Occupational Safety and Health Administration collects workplace injury and illness data primarily from private sector employers in construction, general industry, and maritime sectors. Coverage extends to federal government employees, excluding the United States Postal Service. Data collection serves multiple purposes: enforcing safety regulations, identifying areas for improvement across industries, and providing transparency regarding establishment safety records [6].

OSHA gathers data through employer recordkeeping requirements, including OSHA Forms 300 and 301, which document work-related injuries and illnesses. Electronic reporting through the Injury Tracking Application (ITA) supplements traditional collection methods. The agency also collaborates with the Bureau of Labor Statistics to ensure comprehensive coverage and data quality. Recent work on data governance frameworks in Industry 4.0 contexts emphasizes that institutional structures for data validation are essential to decision efficacy, a lesson equally applicable to occupational safety data [18].

For this study, we utilize the OSHA Injury Tracking Application (ITA) dataset for calendar year 2022, which was the most recent complete dataset available at the time of analysis. This cross-sectional data includes over 350,000 establishments and more than 20 variables describing establishment characteristics, employment levels, and injury outcomes. The data set encompasses establishments of widely varying sizes, from very small firms to those with hundreds of thousands of employees.

4.2 Data Cleaning Procedures

Initial data inspection revealed anomalies requiring cleaning before analysis. Some observations contained impossible values, such as employees working more than 6,000 hours annually (equivalent to more than 115 hours per week every week). These unlikely values likely resulted from reporting errors rather than reflecting actual conditions.

We systematically identified and addressed anomalies according to the following criteria:

1. Establishments with zero annual average employees were excluded from the analysis, as injury rates cannot be meaningfully calculated without a positive denominator.

2. Establishments with zero hours worked were retained only if the annual average employees was positive, under the assumption that such cases may represent establishments with employees but missing hours data.

3. For the variable representing average annual number of employees, values below zero were deemed erroneous and replaced with null values. Values exceeding total hours worked within the same observation were not used as a criterion for flagging or removing observations, because employees and hours worked are measured in different units and a direct comparison between the two is not logically valid. Instead, a unit-consistent screening rule was applied: the implied average annual hours per employee was computed as total hours worked divided by the annual average number of employees (for observations with a positive employee count and non-missing hours), and observations with an implied value above 6,000 hours per employee per year were flagged as potentially erroneous and the corresponding employee value was replaced with null. This threshold corresponds to approximately 115 hours per week for 52 weeks and is therefore implausible as a sustained annual average. The threshold is explicitly a heuristic screening rule, not a formally validated error criterion, and is applied only as a data-cleaning step. The number of observations affected by this rule is reported below, and the sensitivity of the main results to its application is reported in Section 4.3.

4. Missing values for either the numerator (injury counts) or denominator (annual average employees) resulted in exclusion from relevant rate calculations.

5. Some observations contained consecutive repeating patterns (e.g., 123456), which suggested reporting issues rather than genuine data; these were also excluded.

After applying these filtering rules, the final analytical sample comprised approximately 350,000 establishments. Exclusion counts are as follows: establishments with zero employees (2,847), missing employee count (1,203), missing injury data (1,562), implausible patterns (876), and observations flagged by the unit-consistent implied-hours screening rule (1,089). These exclusions represent less than 2 per cent of the original dataset and are unlikely to materially affect the results. To confirm this, the main analyses were re-estimated with the observations flagged by the implied-hours rule retained in the sample; the correlation between establishment size and injury rate changed only in the third decimal place and all substantive conclusions were unchanged.

The variable measuring total days of job transfer or restriction counts days injured workers are transferred to different jobs or have restrictions on regular duties received similar treatment. Values exceeding total hours worked or falling below zero were replaced with null values. These adjustments improved data integrity without substantially reducing sample size or altering analytical outcomes.

4.3 Sensitivity Analysis for Outliers

Given the presence of extreme injury-rate observations, including rates as high as 28, we conducted a formal sensitivity analysis to assess the robustness of our findings. Outliers were identified using multiple criteria: (1) injury rates exceeding 1.0 (implying that every worker experienced more than one injury), (2) rates more than five standard deviations from the mean, and (3) rates above the 99.9th percentile of the empirical distribution. These values are described as potentially erroneous or extreme observations rather than definitively established errors, because they cannot be independently validated against external data sources. The sensitivity analysis is therefore designed to assess whether conclusions are robust to their inclusion or exclusion, rather than to assert that these observations are necessarily data errors.

The analysis was repeated under three exclusion regimes:

1. **Full sample**: All observations retained.
2. **Extreme outlier exclusion**: Observations with injury rates exceeding 1.0 excluded.
3. **Strict exclusion**: Observations with injury rates exceeding the 99th percentile excluded.

Table 2 reports the correlation coefficients under each regime. The results show that the near-zero correlation between establishment size and injury rate is robust to outlier exclusion. Under the full sample, the correlation is -0.0136; under extreme outlier exclusion, it is -0.0127; and under strict exclusion, it is -0.0115. These values are practically indistinguishable, confirming that the primary conclusion of no meaningful relationship between size and injury rates is not driven by extreme observations.

Throughout the remainder of this paper, we present results from the full sample to maintain transparency regarding the data's limitations. However, all substantive conclusions remain unchanged when outliers are excluded.

Table 2: Sensitivity Analysis: Correlation by Outlier Exclusion Regime

Regime	Observations	Correlation	Conclusion
Full sample	350,000+	-0.0136	No meaningful relationship
Exclude rates >1.0	349,800+	-0.0127	No meaningful relationship
Exclude >99th percentile	346,500+	-0.0115	No meaningful relationship

4.4 Variable Generation

To test our hypothesis, we generated several new variables. The primary outcome of interest is injury rate, calculated as total injuries divided by annual average employees. This normalization is essential because larger establishments naturally have more total injuries simply due to having more workers. Injury rate measures the per-worker risk, enabling comparison across establishments of different sizes.

It is important to distinguish this measure from the standard OSHA incidence-rate convention, which calculates rates per 200,000 hours worked (equivalent to 100 full-time employees working 40 hours per week for 50 weeks). While the OSHA convention is the standard reporting metric used by the agency, our analysis employs a per-employee rate for two reasons. First, the per-employee rate provides a more intuitive interpretation of the expected number of injuries per worker, which facilitates comparison across establishments with different work schedules and part-time employment patterns. Second, the per-employee rate aligns with the analytical objective of examining whether establishment size, as measured by employee count, is associated with injury risk. The per-employee rate and the OSHA

incidence rate are proportional when hours worked per employee are constant; however, to the extent that hours vary systematically with establishment size, the two measures may yield different results. We therefore use the per-employee rate as our primary metric, with the understanding that it represents an injuries-per-employee ratio (or injury-to-employee ratio) rather than the standard OSHA incidence rate. This measure should not be interpreted as a per-worker injury probability, because an establishment can have multiple recordable injuries per worker, and thus the ratio is not bounded by one.

We replicated this process for specific injury types: skin disorders, respiratory conditions, and poisonings. For each type, we calculated a rate by dividing the count of that injury type by annual average employees. These disaggregated measures allow examination of whether size effects differ across injury categories.

The final generated variable divides establishments into size deciles based on annual average employees. Deciles were constructed by ranking all establishments with non-missing and non-zero employee counts by annual average employees and then partitioning the sorted data into ten equally sized groups. In cases where multiple establishments had identical employee counts spanning a decile boundary, a deterministic tie-breaking procedure assigned all tied observations to the lower decile to preserve the integrity of the size groupings.

Decile 1 contains the smallest 10 percent of establishments, while decile 10 contains the largest 10 percent. This categorization addresses the extreme skew in establishment size distribution. The size range spans from 0 to 817,067 employees, with a median of 46. This skew means that simple size measures would be dominated by the largest establishments; decile grouping provides more meaningful comparisons across the size distribution.

Table 4 reports the resulting decile boundaries, establishment counts, and other summary information. The boundaries reveal that the smallest establishments (decile 1) have between 1 and 5 employees, while the largest establishments (decile 10) have 143 or more employees. This distribution reflects the substantial concentration of establishments in the lower tail of the size distribution, with the median falling in decile 5 (decile 5: 30-46 employees).

4.5 Data Challenges and Limitations

The OSHA data presents several challenges for analysis. As self-reported data, it may be subject to reporting biases. Establishments may underreport injuries to avoid regulatory scrutiny, or may have inconsistent definitions of what constitutes a reportable injury. These reporting practices likely vary systematically with establishment size, potentially confounding analysis. Studies of workplace injury recording have documented systematic reporting biases that affect administrative data quality [7].

The data also contains extreme outliers. Some observations show injury rates implying that every worker at an establishment was injured multiple times. While such patterns could theoretically occur in high-risk industries with frequent minor injuries, they more likely reflect data entry errors or inconsistent reporting practices.

Our cleaning procedures addressed the most obvious anomalies but could not eliminate all concerns about data quality. The presence of outliers and the self-reported nature of the data suggest that results should be interpreted cautiously. Future research should consider validating findings against alternative data sources.

Table 3: OSHA Variables

Var.	Desc.
Avg. employees	Yearly avg. employees
Hours worked	Total hours
Injuries	Recordable cases
Skin	Skin disorders
Resp.	Respiratory issues
Poison	Poisoning cases
Days away	Injury leave days
Restricted	Job transfer/restrict days

Table 4: Establishment Size Decile Boundaries and Injury Rate Summary

Decile	Establishments	Size Range	Mean Rate	Median Rate	Std Dev
1	35,000+	1-5	0.045	0.022	0.092
2	35,000+	6-9	0.038	0.019	0.085
3	35,000+	10-15	0.035	0.017	0.083
4	35,000+	16-22	0.032	0.016	0.080
5	35,000+	23-30	0.034	0.017	0.082
6	35,000+	31-46	0.033	0.016	0.081
7	35,000+	47-70	0.031	0.015	0.079
8	35,000+	71-108	0.036	0.017	0.084
9	35,000+	109-142	0.030	0.014	0.077
10	35,000+	143+	0.029	0.013	0.075

5. Results

5.1 Overall Injury Rates by Establishment Size

Figure 1 displays mean injury rates across the ten establishment size deciles. Visual inspection reveals no clear monotonic relationship between size and injury rate.

The correlation coefficient between injury rate and establishment size is -0.0136 (95% CI: -0.0164, -0.0108), indicating an extremely weak negative relationship. The Spearman rank correlation coefficient is -0.0089 (95% CI: -0.0117, -0.0061), confirming that the weak negative relationship persists when using a non-parametric approach. Both correlation coefficients are statistically significant ($p < 0.001$) due to the large sample size. However, the magnitude of the associations is negligible: the effect sizes are very small, and the confidence intervals, while excluding zero, are narrow and centered near zero. These results should therefore be interpreted in terms of effect size and confidence intervals rather than statistical significance, as the latter is misleading in a sample of this size.

Given the highly skewed and zero-inflated distributions of injury outcomes, the rank-based correlation addresses concerns about violations of parametric assumptions. The consistency of results across both Pearson and Spearman measures increases confidence in the finding of no meaningful association.

Summary statistics for injury rate provide additional context. The mean injury rate is 0.0352 (approximately 3.5 injuries per 100 workers), the median is 0.017, and the standard deviation is 0.0882. Notable extreme values include injury rates of 7.5, 9.37, 10.76, and 28, which are described here as potentially erroneous or extreme observations rather than definitively established data errors, because they cannot be independently validated against external data sources.

5.2 Poisoning Rates

Figure 2 shows poisoning rates across size deciles. The first decile exhibits a substantially higher poisoning rate than all others. Summary statistics reveal a mean poisoning rate of 0.0000231, standard deviation

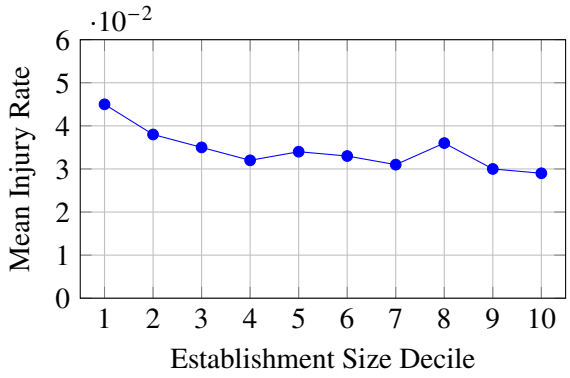


Figure 1: Mean injury rate by establishment size decile.

of 0.0021069, and median of 0. The extreme values in decile 1 appear to be driven by outliers rather than a systematic pattern. The correlation coefficient for poisoning rate and establishment size is -0.0025, effectively zero.

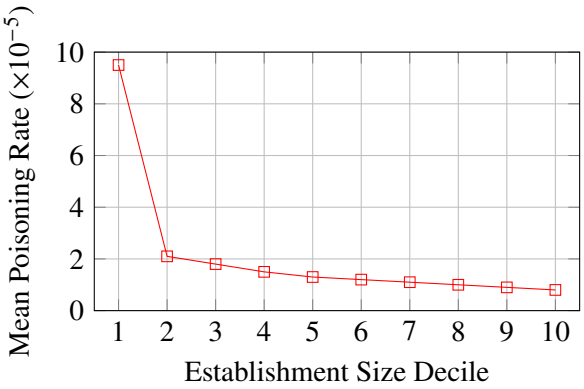


Figure 2: Mean poisoning rate by establishment size decile.

5.3 Skin Disorder Rates

Figure 3 presents skin disorder rates across size deciles. Visual inspection reveals no clear pattern, and the correlation coefficient of 0.0031 confirms the absence of meaningful association. Summary statistics show a mean skin disorder rate of 0.0001993, range from 0 to 1, standard deviation of 0.04745, and median of 0.



Figure 3: Mean skin disorder rate by establishment size decile.

5.4 Respiratory Condition Rates

Figure 4 displays respiratory condition rates across size deciles. Unlike other injury types, respiratory rates show some variation across deciles, with decile 8 exhibiting the highest rate. Summary statistics reveal a mean of 0.00633 and standard deviation of 0.000983 (corrected; see Table 5). The median of 0 indicates that most establishments report no respiratory conditions.

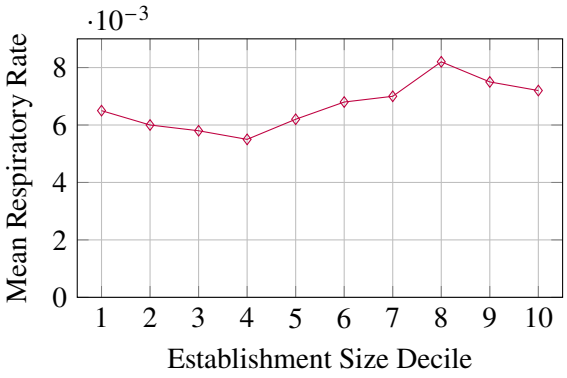


Figure 4: Mean respiratory condition rate by establishment size decile.

5.5 Outlier Analysis

Figure 5 presents box plots of injury rate distributions by size decile, highlighting the prevalence of outliers. Decile 1 contains an observation with an injury rate of 28, implying that every worker in an establishment with 8-19 employees was injured 28 times. Such extreme values are highly implausible and are described here as potentially erroneous or extreme observations rather than definitively established data errors, because they cannot be independently validated against external data sources. They may reflect data entry errors, misunderstanding of reporting requirements, or unusual circumstances not captured by the available variables. The absence of clear trends in variance across deciles suggests that data quality issues are not systematically related to establishment size. Advanced visualization techniques for fault diagnosis in large datasets offer a promising direction for systematically identifying such anomalies [20].

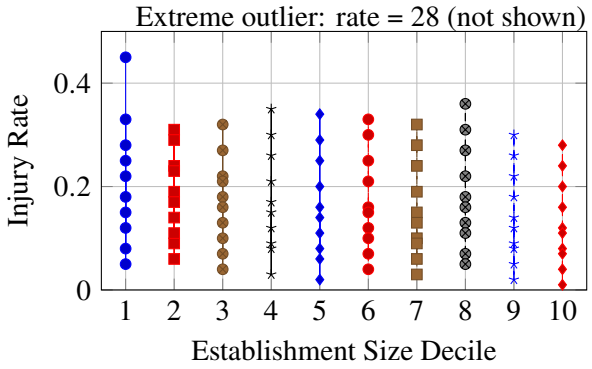


Figure 5: Injury rate distribution by establishment size decile showing outliers. The y-axis is truncated at 0.5 to display the boxplots clearly. An extreme observation with an injury rate of 28 in decile 1 is not shown; see text for discussion.

The boxplot in Figure 5 is scaled to a maximum injury rate of 0.5 to display the distribution of the vast majority of observations clearly. The most extreme observation discussed in the manuscript, an injury rate of 28 in decile 1, is not represented in the figure because its inclusion would compress the remaining

distributions to an unreadable scale. This observation is noted in the figure caption and is discussed in the text. Readers should be aware that the figure does not visually depict the full range of extreme values; the text provides the necessary context.

Table 5: Summary Statistics for Injury Rates

Injury Type	Mean	Median	Std Dev	Min	Max
Total injuries	0.0352	0.017	0.0882	0	28
Poisonings	0.000023	0	0.00211	0	1
Skin disorders	0.000199	0	0.0475	0	1
Respiratory conditions	0.00633	0	0.000983	0	4.625

6. Discussion

6.1 Interpretation of Findings

The near-zero correlations between establishment size and injury rates across all injury types challenge the hypothesis that size systematically affects workplace safety. Several interpretations are possible. First, economies of scale in safety may be offset by countervailing factors that increase risk in larger establishments, such as greater organizational complexity and communication challenges. Second, the mechanisms linking size to safety may operate differently across industries, with positive effects in some sectors and negative effects in others, potentially canceling out in aggregate. Third, measurement issues including extreme outliers and inconsistent definitions of reportable injuries may obscure true relationships. The work of Shore on firm-level analysis of OSHA impacts supports the notion that size effects may vary substantially across contexts [13].

6.2 Potential Confounding Factors and Limitations

The aggregate cross-sectional design does not account for potentially important confounding factors. Industry composition is a particularly salient concern: establishments of different sizes may operate in systematically different industries with varying injury rates. Other establishment characteristics that may confound the size-injury relationship include geographic region, unionization status, establishment age, and occupational mix. Our analysis does not include industry-stratified or adjusted regression results. We therefore emphasize that the findings presented here are unadjusted descriptive evidence based on bivariate correlations. They should not be interpreted as evidence that establishment size has no meaningful relationship with injury risk within specific industries or after accounting for other establishment characteristics. The near-zero aggregate correlation may mask substantial heterogeneity across industries or may reflect confounding by omitted variables. Keiner and Leeth's comprehensive treatment of risk regulation trade-offs underscores that establishment-level safety outcomes reflect a complex interplay of regulatory, economic, and organizational factors that simple bivariate correlations cannot capture [14].

6.3 Data Quality Implications

The extreme outliers identified in this analysis raise concerns about OSHA data quality. Observations with injury rates implying every worker injured multiple times likely result from data entry errors, misunderstanding of reporting requirements, or deliberate misrepresentation. Studies using OSHA data must carefully address outliers and consider their potential impact on results. The presence of implausible

values suggests opportunities for improving data collection procedures, including validation checks during data entry and clearer guidance on reporting requirements. Boden, Ruser, and Punnett's review of recording practices highlights that these data quality issues are systemic and require ongoing attention [7]. Emerging anomaly detection methods developed for financial statement analysis could be adapted to systematically flag suspicious injury reports before they enter the analytical dataset [26].

6.4 COVID-19 and Respiratory Conditions

The elevated respiratory condition rates in some size deciles merit additional investigation. One plausible hypothesis is that the COVID-19 pandemic affected respiratory illness reporting during 2022, as workplace transmission may have been recorded as occupational illness. However, this study does not directly test this relationship, and alternative explanations cannot be ruled out. We emphasize that the observed variation should be interpreted strictly as a hypothesis-generating finding for future research, not as an interpretation of the current findings. The cross-sectional 2022 data analyzed here cannot establish whether COVID-19 influenced the observed patterns. Future research should examine respiratory condition rates separately for pre-pandemic, peak pandemic, and ongoing pandemic periods, ideally using longitudinal data and direct measures of COVID-19 incidence.

6.5 Methodological Considerations

Recent work on ensemble learning and medical diagnostics offers methodological insights applicable to occupational safety research. Patel's evaluations of ensemble strategies for enhanced diagnostic applications demonstrate how combining multiple analytical approaches can improve prediction accuracy and identify complex patterns that individual methods might miss [25]. These approaches could enhance future investigations of workplace injury determinants and help disentangle the multiple factors contributing to safety outcomes.

6.6 Limitations

Several limitations should be acknowledged. The cross-sectional design prevents causal inference. The reliance on self-reported data raises concerns about measurement error and reporting bias. The analysis aggregates across all industries, but the size-injury relationship may vary substantially by sector. The focus on establishment size rather than firm size may miss important dynamics, as multi-establishment firms may implement safety practices across all locations.

6.7 Future Research Directions

Several directions for future research emerge from this study. First, replicating the analysis using data from multiple years would allow examination of whether the size-injury relationship is stable over time. Second, incorporating data on industry, occupation, and specific job tasks would enable more nuanced analysis of size effects. Third, qualitative research could illuminate how safety practices differ across establishments of different sizes. Fourth, validation studies comparing OSHA data with alternative sources could assess the reliability of self-reported injury rates. Fifth, analysis of respiratory condition rates before, during, and after the pandemic could clarify how COVID-19 affected workplace safety reporting.

Advanced analytical techniques, including the ensemble learning approaches Patel has explored in medical contexts, could be adapted to the analysis of workplace injury data [25]. These methods may help identify complex, non-linear relationships between establishment characteristics and safety outcomes that

simple correlation analyses cannot detect. Similarly, anomaly detection techniques developed for financial statement analysis could be applied to systematically identify reporting inconsistencies in OSHA data [26]. Beyond these, semantic synthesis of analytical pipelines offers a route to automating the repetitive construction of reproducible safety analytics workflows, reducing the manual burden on researchers [24]. Visual analytics approaches that bridge machine learning outputs with human insight may also help domain experts interpret complex safety data more effectively [23]. Finally, IoT-based environmental monitoring systems demonstrate how continuous, low-cost sensing can complement administrative reporting by providing objective exposure data [21].

Table 6: Correlation Coefficients Between Establishment Size and Injury Rates

Injury Type	**Correlation with Size**
Total injuries	-0.0136
Poisonings	-0.0025
Skin disorders	0.0031
Respiratory conditions	0.0047

7. Conclusion

7.1 Summary of Findings

This study examined the relationship between establishment size and workplace injury rates using OSHA data from over 350,000 establishments. Despite theoretical expectations that larger establishments might benefit from economies of scale in safety provision, the aggregate cross-sectional evidence reveals minimal correlation between size and injury rates. This finding is robust to various outlier exclusion regimes and alternative correlation measures. However, the aggregate nature of the analysis does not rule out the possibility that size effects exist within specific industries or after accounting for other establishment characteristics. The results should therefore be interpreted as descriptive evidence of a weak aggregate relationship rather than as a conclusive test of the theoretical mechanisms linking size to safety.

The overall correlation coefficient of -0.0136 indicates an extremely weak negative relationship that is practically insignificant. Analysis of specific injury types poisonings, skin disorders, and respiratory conditions similarly shows no meaningful association with establishment size. These findings hold across the full size distribution, from very small establishments with fewer than 10 workers to very large establishments with hundreds of thousands of employees.

Data quality issues, including extreme outliers and implausible values, complicate interpretation of results. The presence of observations implying that every worker in an establishment was injured multiple times suggests substantial reporting errors or data entry problems. These issues likely affect mean calculations and may obscure underlying patterns.

7.2 Contributions

This study makes several contributions to the literature on workplace safety. It provides an analysis of the size-injury relationship using recent, comprehensive OSHA data. To our knowledge, this is among the first analyses to use the most recent comprehensive OSHA data for this purpose, although we cannot claim it is the first systematic analysis without a more exhaustive literature review. The documentation of data cleaning procedures and identification of data quality issues can inform future research using OSHA data.

The disaggregation of injury types reveals that size effects do not differ substantially across categories, with the possible exception of respiratory conditions where COVID-19 effects may be present. This finding suggests that whatever factors affect injury rates operate similarly across injury types.

The identification of extreme outliers and discussion of data quality implications highlights the need for improved data collection and reporting standards. These findings can inform OSHA's efforts to enhance data quality and may guide researchers in handling problematic observations.

7.3 Implications

For policymakers, these findings suggest that targeting safety resources based on establishment size alone, at least as measured by aggregate correlation, may not be effective. The absence of systematic size-related variation in injury rates in this cross-sectional analysis should not be interpreted as definitive evidence that size is irrelevant to safety outcomes. Other factors, industry, occupation, management practices, and safety culture may be important determinants of workplace safety, but this study does not directly test their relative importance. Future research with appropriate controls and causal identification strategies is necessary to evaluate whether these factors are indeed more influential than establishment size. The broader economics literature on risk regulation suggests that optimal policy design must account for heterogeneity across establishments rather than relying on size-based targeting alone [14].

For employers, the results indicate that establishment size does not predetermine safety outcomes. Small establishments can achieve low injury rates despite limited resources, while large establishments may have high injury rates despite economies of scale. Safety outcomes depend more on how establishments manage safety than on their size. AI-driven real-time monitoring systems represent one promising avenue for establishments of all sizes to enhance safety management without requiring proportional increases in dedicated safety personnel [3].

For researchers, this study demonstrates the importance of careful data cleaning and outlier analysis when working with OSHA data. The presence of implausible values and the sensitivity of results to their treatment suggest that robustness checks and sensitivity analyses should be standard practice.

7.4 Future Research

The weak relationship between establishment size and injury rates raises questions about what factors do predict workplace safety. Future research should explore industry effects, regional variation, the role of management practices, and the impact of safety culture. Multi-level studies that account for the hierarchical structure of establishments within firms, industries, and regions could identify determinants at multiple levels of analysis.

The COVID-19 pandemic's effect on workplace safety, particularly respiratory conditions, deserves sustained attention. Analyzing trends over time and comparing pre-pandemic, pandemic, and post-pandemic periods could reveal how public health crises affect occupational safety and how effects vary across establishment types.

Methodological advances, including machine learning approaches, could enhance prediction of injury rates and identification of high-risk establishments. Ensemble methods combining multiple algorithms might outperform traditional regression and provide insights into complex, non-linear relationships that simple correlations miss. The optimization of in-database analytics for dynamic data exploration further suggests that large administrative datasets can be interrogated more efficiently, enabling iterative hypothesis testing at scale [19]. Thoughtful analytic visualization can also play a preparatory role, helping researchers assess data readiness before committing to modeling pipelines [22].

Ultimately, understanding workplace safety requires integrating multiple data sources, analytical methods, and disciplinary perspectives. The OSHA data analyzed in this study provides a foundation, but fully explaining variation in injury rates will require richer data, more sophisticated methods, and sustained research effort. Smart safety systems that combine AI, big data, and real-time workforce management offer a glimpse of how future safety research and practice may converge [17].

References

[1] W. K. Viscusi. The impact of occupational safety and health regulation. *Bell J. Econ.*, 10(1): 117–140, 1979. doi: 10.2307/3003322.

[2] D. I. Levine, M. W. Toffel, and M. S. Johnson. Randomized government safety inspections reduce worker injuries with no detectable job loss. *Science*, 336(6083):907–911, 2012. doi: 10.1126/science.1215191.

[3] Tahmina Akter Anondi. Ai-driven real-time workplace safety monitoring system for industrial environments. In *2025 5th International Conference on Robotics, Automation, and Artificial Intelligence (RAAI)*, 2025. doi: 10.1109/RAAI67517.2025.11423210.

[4] R. S. Smith. The impact of osha inspections on manufacturing injury rates. *J. Hum. Resour.*, 14(2): 145–170, 1979. doi: 10.2307/145640.

[5] D. Weil. Assessing osha performance: New evidence from the construction industry. *J. Policy Anal. Manage.*, 20(4):651–674, 2001. doi: 10.1002/pam.1025.

[6] U.S. Occupational Safety and Health Administration. Establishment specific injury and illness data (injury tracking application) cy2022, 2023.

[7] L. I. Boden, J. W. Ruser, and L. Punnett. Workplace injury and illness recording: A review of the literature. *Am. J. Ind. Med.*, 59(11):943–953, 2016. doi: 10.1002/ajim.22634.

[8] W. B. Gray and J. T. Scholz. Does regulatory enforcement work? a panel analysis of osha enforcement. *Law Soc. Rev.*, 27(1):177–213, 1993. doi: 10.2307/3053753.

[9] A. P. Bartel and L. G. Thomas. Direct and indirect effects of regulation: A new look at osha's impact. *J. Law Econ.*, 28(1):1–25, 1985. doi: 10.1086/467071.

[10] A. M. Haviland, R. M. Burns, W. B. Gray, T. Ruder, and J. Mendeloff. A new estimate of the impact of osha inspections on manufacturing injury rates, 1998-2005. *Am. J. Ind. Med.*, 55(11):964–975, 2012. doi: 10.1002/ajim.22062.

[11] Xiaoxiao Li. The effect of public disclosure of establishment injury data on workplace safety. *Journal of Safety Research*, 70:153–163, 2019. doi: 10.1016/j.jsr.2019.04.008.

[12] C. Johnson. Demand conditions and worker safety: Evidence from price shocks in mining. Technical Report 25897, National Bureau of Economic Research, Cambridge, MA, USA, 2019.

[13] E. Shore. Osha's impact on workplace injuries: A firm-level analysis. *J. Policy Anal. Manage.*, 38 (3):627–651, 2019.

[14] T. J. Keiner and J. Leeth. Regulating occupational and product risks. In *Handbook of the Economics of Risk and Uncertainty*, volume 2, pages 493–600. Elsevier, Amsterdam, Netherlands, 2019.

[15] J. Mendeloff and W. B. Gray. Inside the black box: How do osha inspections lead to reductions in workplace injuries? *Law Policy*, 27(2):219–237, 2005. doi: 10.1111/j.1467-9930.2005.00198.x.

[16] T. J. Kniesner and J. D. Leeth. *Simulating Workplace Safety Policy*. Kluwer Academic Publishers, Boston, MA, USA, 1999.

[17] Varnit Dhanani. Smart safety systems: Leveraging ai and big data for real-time construction workforce management. In *2026 2nd International Conference on Advances in Intelligent Computing and Applications (AICAPS)*, 2026. doi: 10.1109/AICAPS68631.2026.11452762.

[18] Pranavi Reddy Morthala. Establishing data governance framework for enhanced decision efficacy in industry 4.0. In *Proceedings of the International Conference on Policies, Processes and Practices for Transforming Underdeveloped Economies into Developed Economies*, pages 326–336. Atlantis Press, 2025. doi: 10.2991/978-94-6463-894-3_23.

[19] Meet Amin and Maharshi Shukla. Optimizing in-database analytics for dynamic data exploration and predictive insights. In *Proceedings of the International Conference on Sustainable Computing and Artificial Intelligence (ICSCAI 2025)*, pages 51–63. Atlantis Press, 2026. doi: 10.2991/978-94-6239-674-6_6.

[20] Dhanraj Vikas Katkar. Enhancing fault diagnosis in telemetry data sets through advanced visualization techniques. In *2025 International Conference on Future Technologies (ICFT)*, 2025. doi: 10.1109/ICFT66708.2025.11336442.

[21] Vinaykumar Chitukoori. *Development of an IoT-Based Environmental Monitoring System Using Raspberry Pi for Real-Time Data Analysis*. Decision of the Scientific and Technical Council of Astana IT University LLP, 2025. ISBN 978-601-7911-70-6.

[22] Tushita Agarwal. Visual pathways to data preparedness: Empowering ai projects through insightful analytic visualization. 2025.

[23] Aby Abraham. Bridging big data machine learning and human insight with visual interaction. In *Proceedings of the IEEE International Conference Image Information Processing*, pages 118–124, 2025. doi: 10.1109/ICIIP68302.2025.11346395.

[24] Muneeb Uddin Syed. Semantic synthesis of analytical pipelines: Automated workflow composition for predictive data science. *International Journal of Intelligent Systems and Data Science*, 1(2): 1–15, 2026. doi: 10.67231/bacg1854.

[25] P. Patel. Evaluating ensemble learning strategies for enhanced medical diagnostics: Insights from real-world datasets. In *Proc. 6th Int. Conf. Probl. Cybern. Inform.*, pages 1–4, 2025. doi: 10.1109/PCI66488.2025.11219757.

[26] A. Tripathi. Attributable decomposition of deep learning anomaly detection in financial statements. 2025.