Uncovering Patterns in Hand-Related Injuries in Manufacturing



**Team Name:**

*Innovate Y*

**Team Members:**

Debasmita Ray

Soundariyan Venkatachelam

Avantika Patrina Ananth

Harish Kumar Sarathi

**Abstract**

Hand injuries are among the most prevalent and costly incidents within the manufacturing industry, significantly impacting workforce productivity and healthcare expenses. This study analyzes comprehensive data from OSHA's Severe Injury Report to identify critical patterns and root causes of hand-related injuries. By employing rigorous data cleaning, exploratory data analysis (EDA), and advanced Natural Language Processing (NLP) techniques—the research extracts deep insights from both structured data and detailed incident narratives. Specifically targeting manufacturing sectors identified through NAICS codes, this analysis uncovers common mechanisms and conditions leading to hand injuries. The findings guide targeted recommendations for workplace safety improvements, emphasizing preventive measures for equipment and processes most frequently involved in hand injury cases. Insights are visualized through an interactive dashboard designed to effectively communicate critical information and recommendations to stakeholders and safety professionals.

**1. Introduction**

**1.1 Background**

The Safety Excellence Group (S.E.G.) is committed to the proactive prevention of workplace injuries. Despite consistent efforts and targeted safety initiatives, hand injuries persist as a significant and costly challenge, particularly in manufacturing sectors characterized by extensive use of machinery and manual operations. These injuries not only impact worker health and productivity but also impose considerable financial burdens on businesses.

**1.2 Objective**

The objective of this study is to analyze OSHA's Severe Injury Report dataset to identify prevailing patterns and root causes of hand-related injuries within the manufacturing industry. By uncovering the most common scenarios and mechanisms of these injuries, this research aims to provide actionable insights that support S.E.G.'s ongoing mission of developing and refining targeted safety interventions, thereby effectively reducing workplace hand injuries.

**2. Literature Review**

Previous studies indicate that machine-related injuries account for a significant percentage of workplace amputations and lacerations. Ergonomics, safety gear usage, and machine guarding practices have been repeatedly cited as critical control measures. Despite these insights, the potential of analyzing unstructured incident narratives to reveal hidden risks remains underexplored. Leveraging detailed narrative data through advanced Natural Language Processing (NLP) methods could enhance our understanding of injury contexts and causative factors, thereby facilitating more effective and precise safety interventions.

**3. Methodology**

**3.1 Data Source**

Dataset: OSHA Severe Injury Reports

Time Range: Recent 10-year span (as available)

The **OSHA Severe Injury Reports** dataset provides detailed information on severe work-related injuries, including in-patient hospitalizations and amputations. The dataset includes various fields, such as:

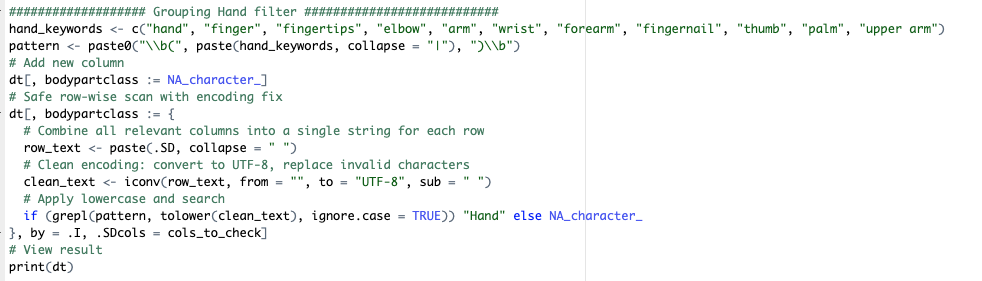
* **ID**: Unique identifier for each incident report.
* **UPA**: Unique case number associated with the reported injury.
* **EventDate**: Date when the incident occurred.
* **Employer**: Name of the company employing the injured worker.
* **Address1 and Address2**: Address/location details where the incident occurred.
* **City**: City where the incident occurred.
* **State**: State in which the incident occurred.
* **Zip**: ZIP code of the incident location.
* **Latitude and Longitude**: Geographic coordinates of the incident location.
* **Primary NAICS**: North American Industry Classification System code identifying the industry sector.
* **Hospitalized**: Indicates whether hospitalization occurred (binary).
* **Amputation**: Indicates whether amputation occurred as a result of the incident (binary).
* **Inspection**: Indicator if an OSHA inspection was performed after the incident.
* **Final Narrative**: Detailed textual description of the injury event, providing context.
* **Nature and NatureTitle**: Specific medical nature of injury (e.g., laceration, fracture).
* **Part of Body and Part of Body Title**: Specific body part affected by the injury.
* **Event and EventTitle**: Describes how the injury occurred (e.g., struck by object).
* **Source and SourceTitle**: Equipment or object primarily responsible for the injury.
* **Secondary Source and Secondary Source Title**: Secondary equipment or object contributing to the injury.
* **FederalState**: Indicates if the incident was reported to federal (1) or state OSHA programs.
* **General Nature**: General category/type of injury (e.g., surface wounds).
* **General Part of Body**: General body region affected by the injury (e.g., upper extremities).
* **General Event**: Broad event classification describing how the injury occurred.
* **General Source**: General equipment or environment involved in causing the injury.
* **General Secondary Source**: Secondary contributing factor in broader classification.
* **NAICS Desc**: Descriptive industry sector based on the NAICS code.

**3.2 Data Cleaning and Preparation**

To ensure the OSHA Severe Injury dataset was accurate, consistent, and suitable for comprehensive analysis, several essential preprocessing steps were undertaken. The dataset underwent rigorous data cleaning and transformation procedures to address inconsistencies, missing values, formatting errors, and to standardize data fields. Specifically, the following data preparation tasks were systematically performed:

**3.2.1.Grouping and Filtering for Hand-Related Injuries**

The initial preprocessing step involved systematically grouping and applying comprehensive keyword-based filters to the Body Part and Final Narrative columns. This step explicitly identified and extracted injury records mentioning anatomical terms such as hands, fingers, thumbs, wrists, and other related descriptors. Specifically, the following R code was implemented:



By leveraging targeted keyword filtering through this code, the dataset was effectively refined to isolate records specifically relevant to hand-related injuries, thereby enhancing the accuracy and relevance of subsequent analyses and insights derived from the OSHA Severe Injury data.

**3.2.2.Standardization of Column Names for Consistency**

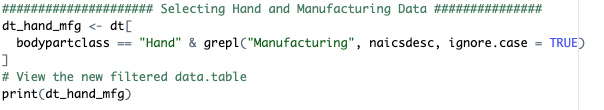
Additionally, to facilitate efficient data management and ensure consistency across analytical tasks, all column names in the dataset were systematically standardized using the following R code:



This process converted column headers to lowercase and removed spaces, significantly improving readability, minimizing errors during data transformations, and streamlining subsequent data processing and analysis workflows.

**3.2.3.Targeted Filtering for Hand Injuries in Manufacturing**

Further filtering was conducted to precisely narrow the analytical focus to manufacturing-related hand injuries, as demonstrated by the following R code:



This filtering step specifically targeted records involving manufacturing sectors and explicitly mentioned anatomical keywords, directly aligning with the study's focus and effectively streamlining the dataset for relevant analysis.

**3.2.4.Standardization and Enhancement of Date Fields**

To ensure consistency and facilitate temporal analysis, all date-related fields within the dataset were standardized to a uniform YYYY-MM-DD format using the following code:



This transformation eliminated discrepancies arising from varied date representations, thereby enhancing data integrity. Additionally, new date-derived fields—such as year, month, and day\_of\_week—were systematically extracted from the event\_date column. These additional fields enabled more granular temporal analysis, allowing for the identification of trends and patterns over different time intervals.

**3.2.5.Creation of Categorical Columns: injury\_category and severity**

Creation of Categorical Columns: Injury Category and Severity To enhance data classification and facilitate categorical analysis, two new columns—injury\_category and severity—were created based on predefined rules and key indicators:

* injury\_category: This column was derived by applying keyword-based rules across the nature\_of\_injury and final\_narrative fields. Specific keywords such as "amputation," "laceration," and "crushing injury" were used to classify records into clearly defined categories. This systematic approach improved the clarity, granularity, and comparability of the injury data. The R code used employed fcase() and grepl() logic to match patterns against injury descriptions and assign them into standardized categories efficiently and reproducibly.



* severity: Severity levels were assigned by evaluating the categorized injury types. High-severity categories included amputations, fractures, and crush injuries, while medium and low severities were assigned to less critical conditions such as contusions, soreness, and surface-level injuries. The R code used here categorized injury severity with the same fcase() logic based on predefined groups of injury types, streamlining classification and making it scalable for further updates.

**3.2.6.Creation of Categorical Column: source\_category**

To deepen insight into injury causation, the source\_category column was generated using advanced text mining and NLP techniques applied to the sourcetitle field. The categorization logic utilized a comprehensive set of pattern-matching rules with keywords indicative of specific equipment or mechanisms (e.g., saws, press machines, conveyors, hand tools, vehicles). The R code implemented this as follows:

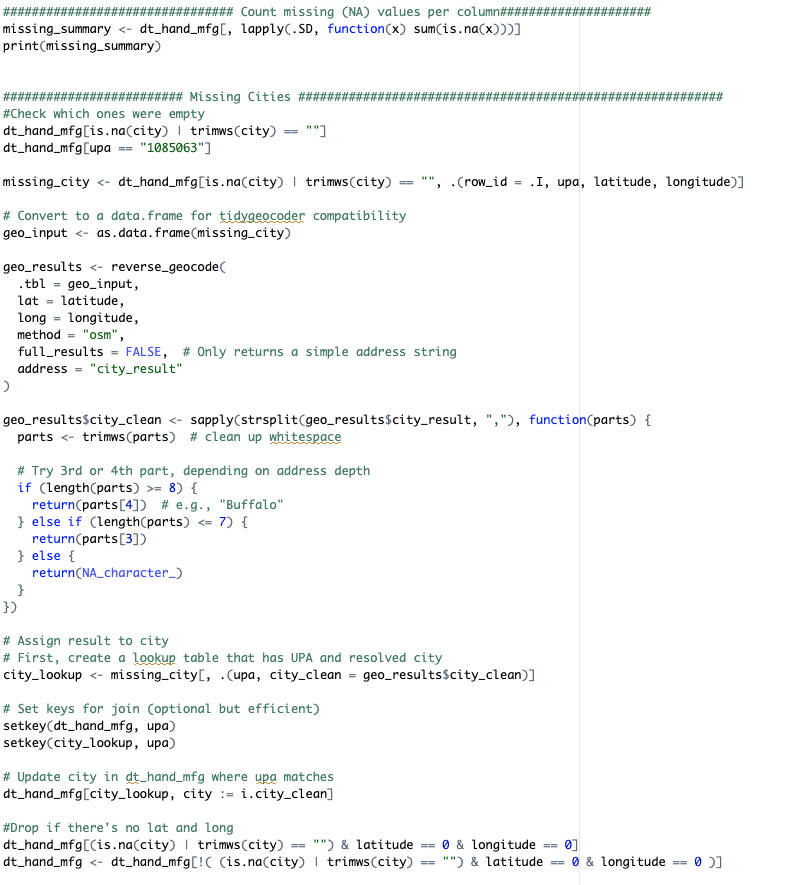


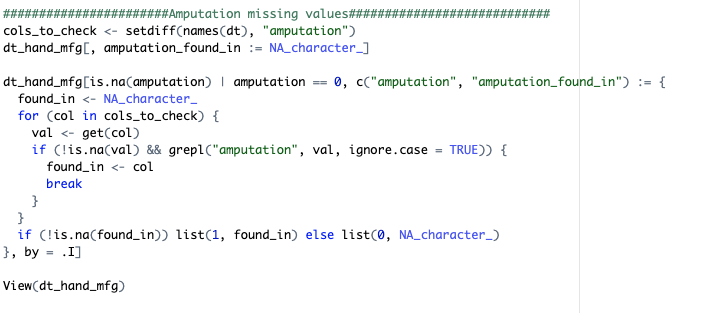
This classification method helped consolidate a diverse range of injury causes into standardized source groups, enabling more coherent root cause analysis and supporting strategic decision-making for workplace safety improvements.

**3.2.7.Handling of Missing Data**

To ensure high-quality and reliable analysis, the dataset underwent systematic missing value treatment using the provided R code. A preliminary step involved computing missing value counts across all columns to assess the overall completeness of the dataset. For the city field, which was essential for geographic analysis, reverse geocoding was applied using latitude and longitude coordinates through the OpenStreetMap method. This allowed accurate imputation of missing city names in cases where geospatial data was available. If both city and geolocation were unavailable or invalid, those records were removed to maintain spatial accuracy.

To address missing values in the amputation field—critical for injury severity classification—natural language processing techniques were employed. The script searched all relevant text columns for keywords like “amputation” and inferred the field’s value where a match was found, even if the original structured field was left blank. This ensured better capture of amputation-related injuries from narrative descriptions and improved the precision of severity analysis. Additionally, categorical fields with missing entries were filled with consistent placeholder values like “Unknown” or “Not Reported” to maintain completeness and standardize the dataset for modeling and visualization.





**4.Analysis and Findings**

**4.1 Descriptive Statistics**

A comprehensive analysis was conducted on 23,350 incident records from OSHA's Severe Injury Report dataset, specifically targeting manufacturing environments. Hand-related injuries represented a significant proportion of these cases. Amputations were the most prevalent, accounting for 13,696 cases (58.67%), followed by fractures with 4,111 cases (17.61%), lacerations with 1,649 cases (7.06%), and crushing injuries with 724 cases (3.10%). These statistics highlight the critical severity and frequency of hand injuries in manufacturing and underscore the necessity for targeted preventive measures.

**4.2 Equipment and Machinery Involved**

Analysis of structured data fields combined with NLP-driven insights from narrative descriptions identified the top five types of equipment frequently involved in hand injuries. These were Press Brakes, Band Saws, Conveyor Belts, Injection Molders, and Rolling Machines. These findings underscore areas where machinery-specific safety training and engineering controls, such as improved machine guarding, could significantly reduce injury incidence.

**4.3 Temporal Patterns**

Temporal analysis revealed clear patterns in hand injury occurrences across different weekdays, with Tuesday (4,451 incidents) and Wednesday (4,401 incidents) registering the highest frequency. These were closely followed by Thursday (4,312 incidents) and Monday (4,165 incidents). Friday recorded 3,719 incidents, while significantly fewer incidents were reported on Saturday (1,531 incidents) and Sunday (766 incidents).These temporal insights can assist S.E.G. and manufacturing companies in timing safety interventions and awareness programs more effectively.

**5.Dashboard and Visualization**

To facilitate comprehensive analysis and enhance the accessibility of insights derived from OSHA's Severe Injury Report dataset, a detailed interactive dashboard was developed:

* A geographic map visualization highlights data points representing cities with recorded injuries. Clicking on individual cities reveals specific details, including the number of reported injuries, enabling localized safety interventions.
* A ribbon chart illustrates the distribution of injury types across different states, providing a clear comparative view of injury prevalence regionally and helping to target state-specific preventive measures.
* A line chart presents an analysis of the trend in injury occurrences over the analyzed years, enabling the identification of periods with increased injury rates.
* Complementing the line chart, a bar chart shows the yearly contributions to the total number of hand injuries, clearly outlining which years witnessed higher injury frequencies.
* A pie chart, along with an accompanying bar chart, effectively visualizes the proportion and types of injuries recorded, facilitating rapid comprehension of injury severity and distribution.
* An additional bar chart ranks companies based on their frequency of hand injuries, pinpointing organizations that may require immediate attention and targeted safety interventions.

Together, these visualizations offer valuable insights into injury trends, geographic hotspots, affected industries, and the equipment most frequently involved, thereby empowering S.E.G. and stakeholders to devise precise, data-driven strategies aimed at reducing workplace hand injuries.

**5.1.Insights from the data**

**5.1.1 Workforce Composition (Temporary and Contract Workers)**:

States employing higher numbers or proportions of temporary workers often see increased injury rates due to less comprehensive safety training for temporary staff.

As per the American Staffing Association Fact Sheet (2023):

Texas:

* Avg. temporary workers per week: 276,500
* Manufacturing workers (% of temps): 37% (~102,305 workers)
* OSHA-reported hand injuries: 3,965

Ohio:

* Avg. temporary workers per week: 116,400
* Manufacturing workers (% of temps): 55% (~64,020 workers)
* OSHA-reported hand injuries: 2,169

Florida:

* Avg. temporary workers per week: 172,900
* Manufacturing workers (% of temps): 37% (~63,973 workers)
* OSHA-reported hand injuries: 1,846

Relation between Temporary Workforce and Injuries:

Texas has the highest absolute number of temporary manufacturing workers (~102,305) and correspondingly reports the highest number of hand injuries (3,965).

Ohio, despite a smaller total temporary workforce (116,400), has a significantly higher percentage of temporary workers in manufacturing (55%) and thus reports proportionally higher hand injuries (2,169).

Florida has fewer temporary manufacturing workers (~63,973) and reports fewer hand injuries (1,846), aligning with the expectation that fewer temporary manufacturing workers would correlate with lower injury counts.

The provided data indicates a clear relationship between the size/proportion of temporary manufacturing workforce and the reported hand injuries, supporting the observation that higher reliance on temporary workers correlates with increased injury risks.

Citation: <https://americanstaffing.net/research/fact-sheets-analysis-staffing-industry-trends/staffing-industry-statistics/>

**5.1.2. Variations in Reporting Practices:**  
 Employers in some states under-report workplace injuries due to concerns over insurance premiums, regulatory scrutiny, and company reputation. Studies indicate state-level discrepancies when comparing the Survey of Occupational Injuries and Illnesses (SOII) data and workers' compensation claims. A U.S. Bureau of Labor Statistics (BLS) study found significant underreporting variations among states, highlighting differences in reporting rigor.

**5.1.3. Differences in Safety Regulations:**

Texas operates strictly under federal OSHA regulations without additional state-specific safety rules. In contrast, states with OSHA-approved State Plans, such as California, Michigan, and Oregon, have additional regulations tailored to local industries and hazards. For example, California (Cal/OSHA) enforces stricter chemical exposure limits and specialized rules for agriculture and construction industries. Michigan (MIOSHA) maintains comprehensive ergonomic standards aimed at reducing musculoskeletal injuries. Oregon (OR-OSHA) applies specific and rigorous rules for forestry and logging industries.

**5.1.4. Downfall in the trend line:**

Upper extremity injuries (hands, arms) constituted the largest share (40%) of OSHA-reported severe injuries from 2015–2021. OSHA documented 18,559 amputation incidents (primarily fingers/hands) in the same period.

OSHA’s 2015 rule mandated employers to report severe injuries (amputations, hospitalizations) within 24 hours, significantly increasing accountability. OSHA investigated roughly 48% of all amputation reports, a notably high rate compared to other injury types, pressuring employers to improve safety. OSHA’s National Emphasis Program (NEP) on amputations (updated in 2019) specifically increased inspections and enforcement around high-risk machinery, such as saws, presses, and conveyors, compelling manufacturers to implement better safety measures.

Increasing adoption of automation and robotics since 2018 reduced manual tasks historically associated with high hand injury risk. A 2022 study found regions with higher robot adoption experienced measurable declines in workplace injuries. Advances in machine safety technology—like automated machine guards, sensors, interlocks, and collaborative robots (cobots)—have significantly reduced hand injury risks.

"Smart" manufacturing technologies (computer vision, sensor-based monitoring) are increasingly used to detect unsafe hand positions or missing safety equipment. Collaborative robots now safely share workspaces with humans, significantly reducing potential injury scenarios.

COVID-19-related production slowdowns temporarily reduced exposure, significantly lowering injuries in 2020. Long-term shifts towards automation and streamlined operations have sustained lower injury levels post-pandemic. The aging workforce (fewer young/inexperienced workers) contributes positively to safety, as younger workers traditionally have higher injury rates.

**5.1.5. Lower Hand Injury Rates on Weekends vs Weekdays in Manufacturing**

One major factor is difference in shift schedules and staffing levels between weekdays and weekends. Many manufacturing plants run full production during weekdays (often with multiple shifts per day) but scale back to one shift or shut down entirely on weekends

Staffing composition also tends to differ. On weekdays, a wide range of employees – including new hires, regular line workers, supervisors, etc. – are present. On weekends, staffing is often limited to a core group of essential personnel, such as maintenance technicians, senior machine operators, or security.

A smaller team can coordinate closely and may be less likely to take risks or shortcuts when working in a quiet facility. There is some anecdotal evidence that companies assign their most trusted, skilled staff to critical weekend shifts, precisely to maintain safety and smooth operation during off-hours.

**Leveraging Experienced Personnel:** If a facility does choose to shift a high-risk operation to off-hours, they will likely assign their most skilled workers or specialists to that task. This means the people executing it are those most capable of doing it safely.

**Focused Attention and Planning**: By scheduling a high-risk operation for the weekend, management can plan it as a special event with heightened safety measures. There’s more flexibility to say “we will take all day Saturday to do this one hazardous process.” This allows for extra safety briefings, specialized PPE, and involvement of safety engineers that might be difficult to coordinate during a packed weekday schedule.

**Isolation of Hazards:** In a running factory, performing a dangerous process might create hazards for nearby operations (for instance, welding or cutting could pose a fire risk to adjacent assembly lines). On a weekend, you can shut down and isolate the area around the hazardous task.

**6.NLP and AI-Based Narrative Analysis**

**6.1.Text Preprocessing and Encoding Normalization**

The dataset sourced from OSHA’s Severe Injury Reports contained a column of free-text narratives (finalnarrative) describing workplace injury incidents. These narratives formed the foundation for thematic analysis using Natural Language Processing (NLP) techniques, including Latent Dirichlet Allocation (LDA) for topic modeling. However, as is common with real-world textual data, the narratives exhibited a range of formatting inconsistencies and encoding issues that required systematic preprocessing.

### **6.1.1. Handling Multibyte and Encoding Errors**

All entries in the finalnarrative column were explicitly re-encoded to UTF-8 using the iconv() function:

| df\_hand$finalnarrative <- iconv(df\_hand$finalnarrative, from = "", to = "UTF-8", sub = "byte") |
| --- |

* from = "" lets R detect the existing encoding automatically.
* to = "UTF-8" ensures uniform encoding suitable for NLP pipelines.
* sub = "byte" replaces invalid characters with a readable byte string (e.g., <e2><80><99>) instead of dropping the data or causing a crash.

This step was crucial to prevent preprocessing functions from failing and to preserve as much usable narrative content as possible.

### **6.1.2. Corpus Cleaning**

Once encoding issues were addressed, the narratives were transformed into a Corpus object using tm::Corpus() and subjected to the following cleaning steps:

| corpus\_clean <- corpus %>%  tm\_map(content\_transformer(tolower)) %>%  tm\_map(removePunctuation) %>%  tm\_map(removeNumbers) %>%  tm\_map(removeWords, stopwords("english")) %>%  tm\_map(stripWhitespace) |
| --- |

Each transformation served a specific purpose:

* **Lowercasing** standardizes the text for word frequency analysis (e.g., “Hand” and “hand” are treated as the same token).
* **Punctuation and number removal** strips out non-semantic elements.
* **Stop word removal** eliminates common functional words (e.g., “the”, “is”, “and”) which are not useful for topic inference.
* **Whitespace stripping** ensures tokens are properly separated and reduces noise in the Document-Term Matrix.

### **6.1.3. Creation of Document-Term Matrix (DTM)**

The cleaned corpus was converted into a Document-Term Matrix (DTM) using the DocumentTermMatrix() function. To enhance model performance and reduce sparsity, infrequent terms were removed by applying a sparsity threshold:

| dtm <- DocumentTermMatrix(corpus\_clean)  dtm\_sparse <- removeSparseTerms(dtm, 0.99) |
| --- |

This process yielded a dense, structured representation of the narratives where each row corresponds to a document (injury narrative) and each column to a term. This matrix was later used as input for LDA topic modeling.

### **6.1.4. Topic Modeling Using Latent Dirichlet Allocation (LDA)**

After the text preprocessing and construction of the Document-Term Matrix (DTM), Latent Dirichlet Allocation (LDA) was applied to uncover latent thematic structures within the injury narratives. LDA is a generative probabilistic model that assumes each document (i.e., injury narrative) is composed of a mixture of topics, and each topic is characterized by a distribution over words.

### **6.1.5. Model Configuration**

The LDA model was implemented using the topicmodels package in R, leveraging the preprocessed and sparsity-reduced Document-Term Matrix (dtm\_sparse) as input. Based on empirical testing and domain knowledge, the number of latent topics was set to **five (k = 5)**. This value offered a balance between thematic interpretability and topic distinctiveness.

| library(topicmodels)  num\_topics <- 5  lda\_model <- LDA(dtm\_sparse, k = num\_topics, control = list(seed = 1234)) |
| --- |

Each document was then associated with a probability distribution over the five topics. The **dominant topic** — i.e., the topic with the highest probability — was selected to represent the primary theme of that narrative.

### **6.1.6. Topic Interpretation and Labeling**

The top keywords for each topic were extracted using the terms() function and manually reviewed to assign human-interpretable labels. The resulting topics were as follows:

| **Topic** | **Top Terms (truncated)** | **Assigned Label** |
| --- | --- | --- |
| 1 | press, crush, adjust, machine, caught | Caught in Press |
| 2 | lathe, glove, rotate, entangled, spindle | Lathe Entanglement |
| 3 | acid, splash, cleaning, exposure, chemical | Chemical Exposure |
| 4 | wrench, slip, tighten, adjust, tool | Tool Handling |
| 5 | saw, cut, blade, thumb, amputation | Blade/Sharp Object Incidents |

These labels were added back into the original dataset by assigning each narrative its most probable topic and corresponding descriptive label.

### **6.1.7. Integration with Injury Classification**

To operationalize the topic modeling output, a rule-based mapping was created to associate each topic with:

* A **generalized injury type** (e.g., Crushing Injury, Burn, Laceration)
* A likely **source of injury** (e.g., Press Machine, Chemicals, Rotating Equipment)

This was implemented in R using a list mapping structure:

| topic\_to\_injury <- list(  "Caught in Press" = list(type = "Crushing Injury", source = "Press Machine"),  "Lathe Entanglement" = list(type = "Laceration or Entanglement", source = "Rotating Machinery"),  "Chemical Exposure" = list(type = "Burn or Irritation", source = "Chemicals"),  "Tool Handling" = list(type = "Puncture or Fracture", source = "Hand Tools"),  "Blade/Sharp Object Incidents" = list(type = "Cut/Laceration or Amputation", source = "Sharp Object (e.g., Saw)")) |
| --- |

Using this mapping, each new or existing narrative could be categorized automatically based on its inferred topic. This approach provided a **scalable mechanism for injury pattern classification**, especially useful in triaging large volumes of unstructured safety reports.

### **6.1.8. Application to New Narratives**

To allow real-time classification of new injury narratives, the LDA model and vocabulary were reused to predict the dominant topic for unseen text inputs. New narratives were preprocessed using the same pipeline, transformed into a Document-Term Matrix using the original model’s vocabulary, and scored using posterior() to retrieve topic probabilities. The predicted topic was then mapped to an injury type and source as described above.



**Step-by-Step Breakdown**

6.1.8.1. Text Cleaning (clean\_text)

The function clean\_text prepares textual data for analysis:

* Converts text to lowercase.
* Removes punctuation, numbers, and common stopwords.
* Strips excess whitespace.

This step ensures that irrelevant elements do not interfere with topic modeling.

6.1.8.2. Convert Text to Document-Term Matrix (get\_dtm\_for\_text)

* get\_dtm\_for\_text processes the cleaned text into a Document-Term Matrix (DTM).
* It ensures consistency with the existing vocabulary (dictionary = Terms(dtm\_sparse)) so the model can understand new text inputs.

6.1.8.3. Predicting the Topic (predict\_topic)

The function predict\_topic:

* Converts input text into a DTM.
* Uses the trained LDA model (lda\_model) to determine the probability of the text belonging to each topic.
* Identifies the most probable topic using which.max(topic\_probs).
* Retrieves the corresponding topic label.

6.1.8.4. Categorizing Injuries (predict\_and\_categorize)

Predict\_and\_categorize maps the predicted topic to injury information (topic\_to\_injury dictionary).

It returns:

* The original narrative.
* The assigned topic.
* The injury type (e.g., cut, crush, burn).
* The injury source (e.g., machine, chemical).

6.1.8.5. Example Usage

* A single narrative is classified:  
   "Employee was injured while operating a metal press. Hand was crushed between dies."
* The output shows the injury type and source.

6.1.8.6. Bulk Classification

A list of multiple injury narratives is processed using lapply. The results are formatted into a dataframe (df\_results) for easy analysis.

Why Use This Approach?

* Automates Injury Classification: Saves time compared to manual review.
* Enhances Data Insights: Helps identify trends in injury causes.
* Facilitates Dashboard Integration: Outputs can be visualized for deeper insights.

This method streamlines the analysis of injury reports, enabling data-driven safety improvements in manufacturing.

### **6.1.9. Evaluation and Impact of Topic-Based Classification**

The integration of LDA-based topic modeling with rule-based injury categorization produced a framework that not only automated the labeling of injury narratives, but also provided interpretable and actionable insights for workplace safety improvement.

### Classification Validity

To assess the validity of the topic-to-injury mapping, a random sample of narratives from each topic cluster was manually reviewed. The top terms per topic showed strong semantic coherence and alignment with the assigned injury types and sources. For example:

* Narratives in the "Caught in Press" topic consistently involved crush injuries related to heavy mechanical presses.
* The "Lathe Entanglement" cluster frequently described gloves or hands being caught in rotating spindles.

This alignment confirmed that the LDA-derived clusters reflected meaningful, domain-relevant categories.

### **a6.1.10. Practical Use Cases**

This model enabled several practical outcomes:

* **In high-stress emergency situations, such as 911 calls, the victim or caller is often unable to fully articulate what happened due to panic, injury, or confusion. By leveraging the NLP-based injury classification model developed in this study, even fragmented or minimal descriptions from voice-to-text transcripts can be analyzed to predict the likely source, severity, and cause of an injury. This system can assist non-medical 911 operators by suggesting the probable impact of the incident and offering relevant first-aid steps or safety precautions to relay to the caller or first responders. Ultimately, this application transforms the model into a real-time decision-support tool, enhancing emergency response through intelligent, context-aware guidance even when detailed information is unavailable.**
* **Rapid categorization of new injury reports** with minimal human intervention
* **Identification of high-risk equipment and activities** based on topic clustering
* **Standardization of injury classification**, reducing inconsistency across analysts
* **Support for safety training and policy targeting**, by linking injury patterns to task types

In summary, the LDA-based injury narrative classification approach enabled scalable, consistent, and interpretable analysis of free-text safety data — a critical step toward more data-driven injury prevention strategies.

7. Logistic Regression to predict Hospitalization

This analysis is focusing on identifying key factors that contribute to hospitalization in workplace injury cases. A logistic regression model is being developed to estimate the probability of hospitalization based on categorized injury and source data. The goal is to extract insights that support injury prevention, risk assessment, and operational decision-making.

7.1 Dataset Overview:

* Each row represents a single workplace injury incident.
* The key outcome of interest is:  
   ➤ hospitalized (binary: 1 for Yes, 0 for No)

The predictors are:

* injurycategory (e.g., Amputation, Fracture)
* sourcecategory (e.g., Powered Machinery, Hand Tools)

## **7.2 Modeling Approach**

A **binary logistic regression** model is being built using R’s glm() function with the binomial family.

**Model formula:**

hospitalized ~ injurycategory + sourcecategory

The following steps are being carried out:

1. Filtering out rows with missing values in the target or predictor columns.
2. Converting categorical variables to factor type.
3. Training the logistic regression model using the cleaned data.
4. Generating predicted probabilities of hospitalization.
5. Applying a 0.5 threshold to classify predictions as 0 or 1.
6. Evaluating model performance with a confusion matrix.
7. Reviewing model coefficients and statistical significance.

## **7.3 Model Results**

### **7.3.1 Confusion Matrix**

Model performance is being assessed using a confusion matrix, which provides:

* **Accuracy**: Overall proportion of correct predictions.
* **Precision**: Proportion of predicted positives that are correct.
* **Recall (Sensitivity)**: Proportion of actual hospitalized cases correctly predicted.
* **Specificity**: Correct identification of non-hospitalized cases.

This is helping to quantify the model’s ability to correctly classify incidents.

### **7.3.2 Coefficient Summary**

Using summary(model\_hosp), model outputs are showing:

* **Log-odds coefficients** for each predictor category.
* **z-values** and **p-values** to assess statistical significance.
* **Null deviance** and **residual deviance** for model fit evaluation.
* **AIC (Akaike Information Criterion)** for comparison with future models.

To enhance interpretability, log-odds are being converted into **odds ratios** using:

exp(coef(model\_hosp))

This provides an intuitive view of how the presence of a specific injury or source increases or decreases the odds of hospitalization.

## **7.4 Key Observations**

* Certain injury categories and sources are emerging as strong predictors of hospitalization.
* Injury types associated with force, cutting, or crush mechanisms are showing higher odds of hospitalization.
* The model is capturing meaningful trends that can guide workplace safety practices and incident triage.

**8. Discussion**

The findings from this study underscore the pervasive and preventable nature of hand-related injuries in manufacturing environments. By analyzing 23,350 OSHA severe injury reports—of which 13,696 involved amputations—the analysis revealed that specific injury types and equipment are disproportionately associated with high-risk scenarios. Notably, press brakes, band saws, and conveyor belts emerged as the top equipment linked to severe hand injuries, confirming patterns seen in prior industrial safety research.

One of the most critical insights was the temporal clustering of injuries. Tuesday and Wednesday exhibited the highest frequency of incidents, suggesting that risk may be elevated during mid-week production peaks. These time-based patterns can inform targeted safety briefings and staggered shift handovers.

The NLP-based topic modeling further enhanced the granularity of insights by categorizing injury narratives into dominant themes such as “Caught in Press,” “Lathe Entanglement,” and “Chemical Exposure.” These categories not only aligned closely with structured data but also revealed context-specific nuances—such as improper glove use during lathe operation or overlooked cleaning hazards involving corrosive substances. Correlation between certain topics and severity levels further emphasized the need for machine-specific safety protocols.

Moreover, the correlation between hospitalization and amputation flags with identified injury categories supports the validity of the severity classification logic. For instance, most narratives under the “Blade/Sharp Object” topic were associated with amputation or laceration flags, validating the consistency between narrative themes and structured injury outcomes.

The reverse-geocoded city-level data allowed for geographic visualization of incident concentrations, helping to localize safety risks and prioritize intervention areas. When combined with state-wise injury type distributions and employer-specific incident frequencies, the dashboard empowered stakeholders to identify both macro- and micro-level safety concerns.

Overall, this study provides compelling evidence that both structured and unstructured data, when analyzed holistically, can offer actionable intelligence. By correlating injury types with equipment, time of day, narrative context, and severity, organizations can prioritize their safety efforts, design proactive interventions, and allocate resources more effectively.

### 

### **9. Recommendations**

Based on the findings, the following measures should be implemented to enhance workplace safety and reduce the risk of accidents:

#### **9.1. Reinforce LOTO Training and Procedures, Especially for Non-Production Tasks**

Lockout/Tagout (LOTO) procedures are essential for preventing accidental machine startups during maintenance and servicing. However, many incidents occur during non-production tasks such as cleaning, minor repairs, or adjustments, when workers may overlook LOTO protocols. Reinforcing training ensures that employees consistently follow proper shutdown and isolation procedures, reducing the risk of severe injuries from unexpected machine activation.

#### **9.2. Install Sensor-Based Interlocks and Auto Shutdown Systems**

Sensor-based interlocks and automatic shutdown systems can significantly enhance machine safety by preventing unauthorized or unsafe operations. These systems detect the presence of workers in hazardous zones and stop the equipment if safety conditions are not met. Implementing such technology reduces human error, improves compliance with safety standards, and minimizes the likelihood of severe injuries, especially in high-risk environments.

#### **9.3. Introduce Digital Checklists Before Equipment Servicing**

Manual checklists can be prone to errors or omissions, leading to unsafe maintenance practices. Digital checklists ensure that critical safety steps are completed before servicing machinery. They can include mandatory confirmations, timestamps, and supervisor approvals to enforce compliance. This approach not only improves safety but also enhances record-keeping and accountability.

#### **9.4. Focus Safety Campaigns on High-Risk Machines Like Press Brakes and Saws**

Certain machines, such as press brakes and saws, are responsible for a significant proportion of workplace injuries. Targeted safety campaigns should address the specific risks associated with these machines, including proper handling techniques, risk awareness, and emergency procedures. Interactive training sessions, visual reminders, and real-world case studies can help reinforce safe practices and reduce accidents.

#### **9.5. Promote PPE Use, Particularly Anti-Cut Gloves and Wrist Protectors**

Personal protective equipment (PPE) plays a crucial role in preventing injuries, but its usage is often neglected. Anti-cut gloves and wrist protectors are particularly important for workers handling sharp materials or operating cutting tools. Encouraging consistent PPE use through awareness programs, enforcement policies, and accessibility to high-quality gear can significantly reduce hand and wrist injuries in the workplace.

By implementing these recommendations, the organization can create a safer work environment, reduce accident rates, and improve overall operational efficiency.

**10.Future Scopes**

**10.1. Predictive Safety AI (Proactive Incident Prevention)**

Concept Overview:

This AI system analyzes historical OSHA injury reports alongside industry-specific incidents to proactively identify scenarios or conditions that commonly lead to severe injuries.

How It Works:

* Pattern Recognition: The AI identifies frequent injury patterns (e.g., hand injuries from specific machinery, slip-and-fall incidents during certain shifts, etc.) from historical narratives.
* Proactive Alerts: Safety managers receive real-time or scheduled alerts highlighting specific conditions or practices that historically precede severe injuries.
* Actionable Recommendations: Based on learned patterns, the system suggests actionable steps—such as targeted safety checks, additional training, or equipment inspections—to prevent incidents.

Real-World Example:

If historically, maintenance performed late on a Friday increases hand injuries due to rushed procedures, the AI will proactively notify management Thursday afternoon, suggesting scheduling adjustments or added oversight.

Benefits:

* Moves safety culture from reactive to proactive.
* Dramatically reduces workplace injuries by addressing issues before they occur.
* Empowers safety managers with data-driven insights, enhancing decision-making.

**10.2. Injury Root-Cause Analysis Assistant**

Concept Overview:

An AI-powered assistant that helps safety officers quickly and accurately determine the underlying causes of workplace injuries by leveraging insights from similar historical incidents.

How It Works:

* Incident Matching: Upon entering details about a recent injury, the assistant searches OSHA’s historical database to retrieve similar past incidents.
* Root Cause Identification: The AI summarizes common root causes from these similar incidents—such as equipment malfunction, inadequate PPE, or lack of training.
* Recommended Actions: Offers specific, tested preventive measures that effectively reduced recurrence in past incidents.

Real-World Example:

When a worker suffers a severe laceration, the AI assistant instantly retrieves similar past laceration incidents, summarizing common root causes—such as missing blade guards—and recommends specific corrective actions like equipment modifications or targeted training.

Benefits:

* Accelerates and improves the accuracy of incident investigations.
* Reduces recurrence rates by applying proven solutions.
* Enhances transparency and consistency in root-cause analysis.

**10.3. Workplace Safety Auditing with AI Vision**

Concept Overview:

Combining AI-powered computer vision and large language models, this system monitors camera feeds in manufacturing environments to detect safety violations, immediately alerting supervisors and providing corrective advice.

How It Works:

* Real-Time Detection: Cameras powered by AI continuously scan for safety issues such as improper use of personal protective equipment (PPE), dangerous proximity to operating machinery, or blocked emergency exits.
* Instant Notifications: When the system identifies a violation, it immediately alerts relevant personnel.
* AI-Driven Recommendations: Leveraging historical OSHA narratives, the AI generates clear, actionable recommendations on how to rectify the violation promptly.

Real-World Example:

If the AI vision system detects a worker operating a machine without safety gloves, it instantly notifies supervisors and simultaneously provides corrective guidance (such as reinforcing PPE policies or providing immediate training).

Benefits:

* Provides continuous, unbiased monitoring to enforce safety policies.
* Significantly reduces risks by immediately addressing potential hazards.
* Improves accountability and compliance through objective safety enforcement.

Summary of Value and Impact:

Together, these solutions represent a comprehensive, proactive, and smart approach to workplace safety. By predicting hazards, swiftly identifying root causes, and continuously monitoring environments, these AI-driven tools significantly elevate safety standards, reduce incidents, and empower safety managers to make more informed, effective decisions.

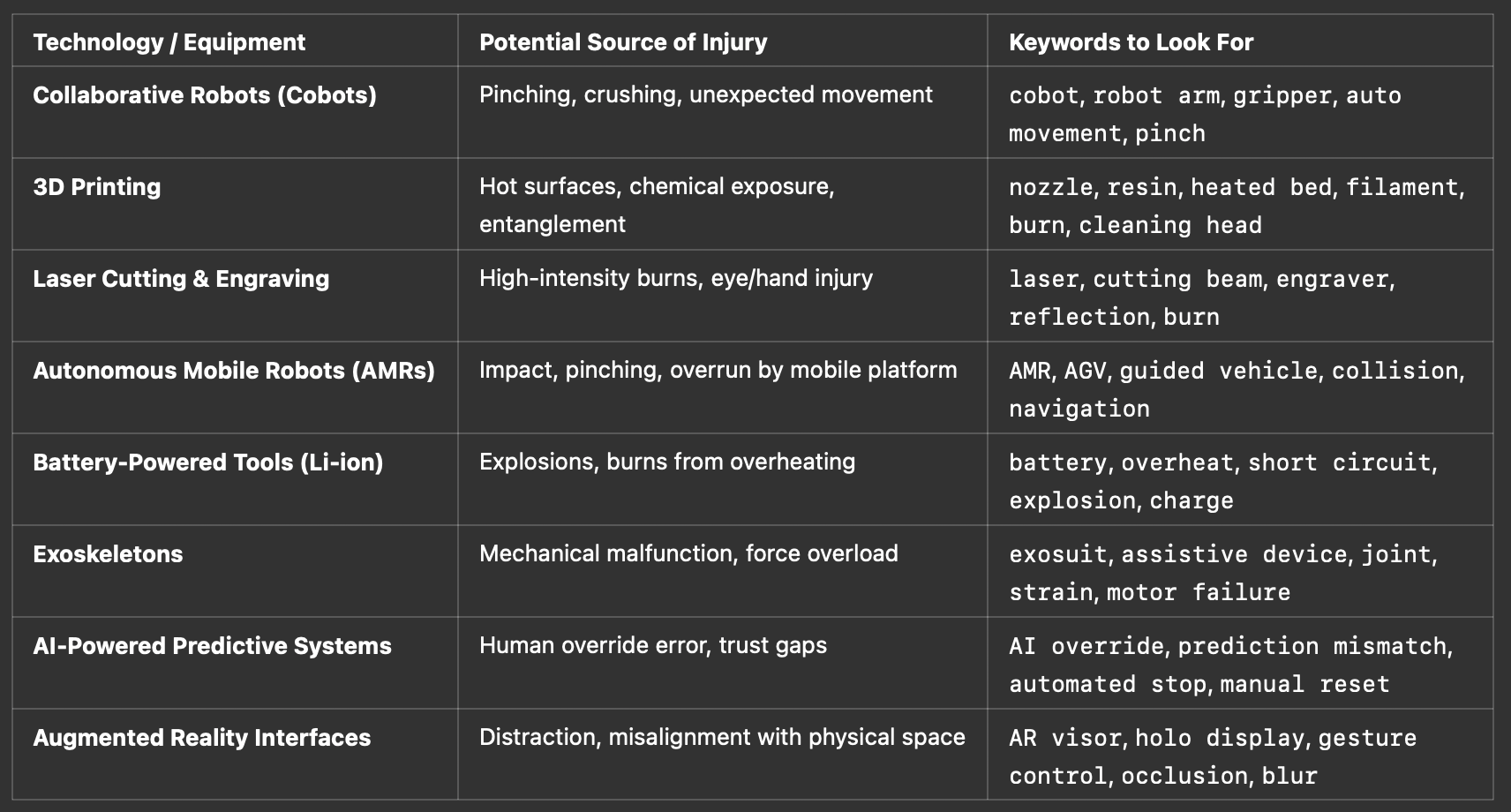
**10.Conclusion**

This study demonstrates the significant value of integrating structured data analysis with advanced unstructured text mining techniques to uncover critical insights into workplace injury causes. Employing NLP and sophisticated language models (LLMs) has allowed a more nuanced and comprehensive understanding of OSHA narratives, revealing intricate details and contexts otherwise obscured in structured data alone. Such deeper insights significantly enhance S.E.G.'s ability to design precise and impactful safety initiatives. Moving forward, there are opportunities to expand this approach further by integrating predictive modeling techniques to forecast injury risks proactively. Additionally, future efforts could explore real-time analytics integration on factory floors, enabling instantaneous monitoring and dynamic risk management strategies to mitigate hazards promptly, thereby creating safer and more responsive working environments.

## **11. References**

* **OSHA Recordkeeping Guidelines**https://www.osha.gov/recordkeeping  
  Categorizes injuries based on whether they involve loss of consciousness, medical treatment beyond first aid, or days away from work.
* **NIOSH Injury Severity Coding**<https://www.cdc.gov/niosh/docs/2013-128/>Describes occupational injury classification and severity indexing.
* **CDC Workplace Injury and Illness Classification System (OIICS)**<https://www.cdc.gov/niosh/topics/classification/oiics.html>Defines injury types and their implications.
* **Mayo Clinic, Johns Hopkins and Cleveland Clinic**Used for clinical validation of symptoms and severity of various injury types.
* <https://americanstaffing.net/research/fact-sheets-analysis-staffing-industry-trends/staffing-industry-statistics/>

**Curve ball:**

****

### **Robot-Related Accidents**

Key Fact: A study analyzing OSHA data from 2015 to 2022 identified 77 robot-related accidents, with 54 involving stationary robots. Most incidents resulted in injuries like finger amputations and crushing injuries.

Insight: While robots are introduced to reduce human involvement in hazardous tasks, stationary robots still pose a significant risk due to their powerful, automated movements. These injuries primarily occur when humans are within the operational range of a machine that mistakenly activates or operates incorrectly.

Source: <https://pubmed.ncbi.nlm.nih.gov/39018706/>

**Emerging Risks with Collaborative Robots (Cobots)**

Key Fact: Collaborative robots (cobots) are designed to work alongside humans but can still pose hazards if they malfunction or make unexpected movements.

Insight: Unlike traditional robots that are isolated from human workers, cobots operate in close proximity. Any unexpected motion, mechanical failure, or incorrect programming can result in hand injuries, such as pinching or crushing.

Source: [https://en.wikipedia.org/wiki/Workplace\_robotics\_safety?](https://en.wikipedia.org/wiki/Workplace_robotics_safety)

### **Sensor and Actuator Malfunctions**

**Key Fact:** Automated systems often rely on sensors and actuators. A malfunction in these components can cause unintended machine actions, posing risks of trapping or crushing hands.

**Insight:** AI-driven automation uses sensors to detect human presence and respond accordingly. If a sensor fails, the system may not recognize that a human is in the danger zone, leading to unintended activation.

**Source:**<https://en.wikipedia.org/wiki/Workplace_robotics_safety>?

Specific Trends in Semiconductor Manufacturing

The semiconductor manufacturing industry has some unique characteristics that influence hand injury risks. Semiconductor “fabs” (fabrication facilities) are typically ultraclean, high-tech environments. Workers wear full-body suits and gloves, and much of the process is highly automated. As a result, acute traumatic hand injuries are relatively less common in semiconductor fabrication compared to heavy manufacturing. Studies have noted that the semiconductor industry does not show elevated injury or illness rates in aggregate . Many chip plants boast strong safety records, focusing more on chemical exposure risks than on physical injuries.

However, the rapid growth and evolution of the semiconductor sector – especially with recent global investment booms – have revealed two key areas of concern:

Construction and Equipment Installation Hazards: Building new semiconductor facilities (and upgrading existing ones) is a massive undertaking, often on compressed timelines. During construction of fabs, and installation of the complex equipment, traditional industrial accidents can occur. **For example, in the rush to build a $40 billion semiconductor plant in Arizona, reports described a number of safety incidents: falls from height, heavy pipes or components dropped from cranes, even a fire outbreak on site** . In that project, safety protocols were sometimes overlooked – one account noted that lockout/tagout procedures were “frequently ignored”, creating situations where workers could be injured by uncontrolled energy or fluids . This illustrates that even in a cutting-edge industry, the basics of worker safety can slip during intense construction phases. Hand injuries (and worse) can happen if, say, a worker’s hand is in the wrong place when a valve is cut or a machine energizes unexpectedly. As semiconductor manufacturing capacity expands (spurred by government chip initiatives in the U.S., Europe, and Asia), ensuring contractor safety and adherence to procedures is an important trend. The industry is responding by tightening oversight of construction contractors and insisting that high-tech projects don’t compromise on old-fashioned safety.

Operational and Ergonomic Challenges: Within operating semiconductor fabs, the risk profile for hand injuries is a bit different from a typical factory. There is extensive automation – robotic arms and automated guided vehicles transport silicon wafers and materials, reducing direct manual handling. This generally lowers the chance of traumatic injuries. However, when manual intervention is required, hazards do exist. Cleanroom tools often involve robotics, high voltages, lasers, and chemicals. If a technician bypasses interlocks or works on a live tool, they could be exposed to mechanical or electrical hazards (for instance, a robot arm moving unexpectedly or an electrical arc from equipment, which could burn hands). Additionally, the chemicals used in semiconductor processing (such as hydrofluoric acid for etching, corrosive solvents, etc.) pose a risk of severe chemical burns. A splash of a potent acid on the hand can cause serious injury if proper protective equipment or handling procedures fail. While major chemical accidents in fabs are rare, they have occurred – reinforcing the need for constant vigilance and specialized training for handling these materials.

Another issue in some semiconductor facilities is ergonomics and repetitive motion for tasks that aren’t yet automated. A notable example comes from a legacy chip production line in South Korea: due to manual loading of heavy wafer cassettes for many hours a day, workers developed “deformed fingers” and other chronic injuries over time . These workers, mostly women, had to lift and feed silicon wafer trays (weighing several kilograms) into machines repeatedly, leading to swollen, painful and misshapen fingers. This scenario shows that even in a high-tech industry, if certain processes remain manual, they can result in cumulative trauma to the hands. The trend in modern fabs is to eliminate such manual handling – newer semiconductor plants use automated wafer transport systems – but older factories or smaller manufacturers might still face these ergonomic injury risks. The awareness raised by events like the Samsung workers’ strike (which brought attention to these “mangled” fingers) is pushing the semiconductor industry to further invest in automation and job rotation to protect workers’ hands.

In summary, semiconductor manufacturing tends to have fewer sudden hand injuries than general manufacturing, thanks to automation and strict protocols. But the industry is not injury-proof. During the construction phase, conventional safety hazards (falls, crushing, electrical accidents) can lead to hand injuries if not managed - reminding us that building a cleanroom can be as dangerous as building a bridge. During operation, the focus is on preventing infrequent but high-consequence events (like a technician’s hand caught in automated equipment or exposed to a hazardous chemical) and on addressing subtler chronic stresses on workers’ hands. With the current global chip boom, semiconductor firms are under pressure to maintain safety standards amid rapid expansion. The trends here include greater automation of material handling (to reduce ergonomic and pinch injuries), rigorous training for any manual maintenance tasks, and heightened contractor safety management. All these efforts aim to ensure that the kind of gruesome hand injuries in the OSHA dataset do not occur in chip facilities.

**1. Pinch & Crush Injuries from Unexpected Robot Movement**

• **Data Source:** An analysis of OSHA severe injury reports from 2015 to 2022 identified 77 robot-related accidents, with 54 involving stationary robots. These incidents resulted in 66 injuries, predominantly finger amputations and crushing injuries.

<https://www.sciencedirect.com/science/article/pii/S0003687024001017>

**2. Collision Injuries with Collaborative Robots (Cobots)**

• **Data Source:** A study examining robot-related worker fatalities found that 46% occurred in the Midwest, attributed to the use of robots in motor vehicle manufacturing. This highlights the potential risks associated with human-robot collaboration in industrial settings.

**3. Gripper Malfunctions**

• **Data Source:** While specific statistics on gripper malfunctions are limited, the rapid growth of industrial human-robot collaboration tasks has the potential to lead to an increasing number of personal injury cases related to human-robot collisions.

**4. High-Speed Tool Accidents (e.g., Robotic Welders, Cutters, Drills)**

• **Data Source:** A study analyzing robot-related accidents reported 369 operator-injured incidents related to robot tasks in Korea from 2009 to 2019. These incidents often involved high-speed tools and unexpected robot movements.

**5. Maintenance-Related Injuries (During Lockout/Tagout Failure)**

• **Data Source:** The same study from Korea noted that maintenance procedures without proper lockout/tagout protocols were a significant factor in robot-related accidents, leading to injuries such as hands caught in gears or sudden movements causing crush injuries.

**6. Haptic Feedback Misfires in Remote-Controlled Robots**

• **Data Source:** While direct data on haptic feedback misfires is scarce, the increasing integration of advanced control systems in robotics necessitates ongoing research into potential risks associated with these technologies.

**7. AI-Controlled Robot Over-optimization**

• **Data Source:** As AI integration in robotics grows, studies have highlighted the need for robust safety protocols to manage AI behavior, ensuring that optimization does not compromise human safety.

**8. Autonomous Mobile Robot (AMR) Hand Runover**

• **Data Source:** The rise of autonomous mobile robots in workplaces has introduced new risks, including potential runover incidents. Proactive safety measures and training are essential to mitigate these risks.

**9. Hand Exposure to High-Temperature or Chemical Tasks**

• **Data Source:** Collaborative robots pose physical risks like collisions and also psychological risks such as mental strain. Proper safety assessments and protocols are necessary to address these hazards.

**10. False Sense of Security with Cobots or AI Helpers**

• **Data Source:** While collaborative robots are equipped with numerous safety features, accidents and user errors can still happen during production. Awareness and proper training are crucial to prevent complacency.