



Occupational disease compensation claims in Italy: an epidemiological analysis of national insurance data with predictive insights

MARTINA PADOVAN¹, ALESSANDRO PALLA²

¹ Preventive Medicine, Tuscany North-West Health Local Unit, Italy; ² Intel Corporation, Santa Clara, CA 95054, USA

Keywords

Occupational diseases • Epidemiological surveillance • Impairment severity • Neural networks • Predictive modelling

Summary

Introduction. Occupational diseases remain an underreported public health problem. In Italy, national insurance claims provide a useful view of recognized occupational diseases, although they do not measure the full burden of disease in the workforce. Few studies have combined descriptive analysis of insurance claims with predictive modelling to assess factors associated with recognized impairment.

Methods. We conducted a retrospective observational analysis of 310,189 occupational disease claims from the Italian National Institute for Insurance against Accidents at Work (INAIL) open data, covering the period 2020 to 2024. Multivariable linear regression was used to estimate associations between demographic, clinical, and occupational factors and the degree of impairment. Neural-network models with embedding layers were developed to classify claims with recognized impairment (>0%) and to predict impairment severity.

Results. Musculoskeletal disorders represented the largest share

of claims, whereas neoplasms and asbestos-related diseases were associated with the highest impairment. Diagnostic category and causal agent were the strongest predictors of severity among claims. The classification model achieved 98.15% accuracy and 99.99% sensitivity for recognized impairment. The regression model attained a mean absolute error of 2.31% and $R^2=0.71$, outperforming the linear regression baseline $R^2=0.55$.

Conclusion. Combining descriptive epidemiology with machine learning yields actionable insights for prioritizing high-risk claims and supporting administrative decision-making in occupational health surveillance. These findings represent a proof of concept given inherent administrative data limitations, including underreporting, lack of workforce denominators, and potential data bias. Future developments should incorporate explainable artificial intelligence techniques to bridge the gap between predictive performance and clinical interpretability.

Introduction

Occupational diseases represent a massive component of the global work-related disease burden, with recent estimates attributing approximately 2.58 million deaths to work-related illnesses in 2019 [1, 2]. These pathologies are linked to a wide range of risk factors, including exposure to carcinogens, ergonomic stressors, noise, and long working hours, the latter being a leading risk factor for ischemic heart disease and stroke. Beyond the human cost, the economic impact is profound: the total loss resulting from work-related disability and death is estimated at 6% of the global GDP [3].

Despite these alarming figures, the true scale of the problem remains largely hidden due to the “iceberg phenomenon”, where official records often represent only the visible portion of the total burden. Global systematic reviews and conceptual models indicate that underreporting rates are exceptionally high, ranging between 50% and 95%, driven by “conceptual filters” such as job insecurity, limited worker knowledge of legal benefits, and diagnostic difficulties [4, 5].

In the European context, the European Agency for Safety and Health at Work (EU-OSHA) defines occupational

diseases as those for which a causal link is established between workplace exposure and the pathology [6]. Current European strategies emphasize the need for sustainable work and improved conditions to address the evolving nature of risk in a changing labor market [7].

To address the need for comparative data, the European Occupational Diseases Statistics (EODS) provide experimental statistics based on cases recognized by national authorities [8]. However, interpreting regional data remains complicated by significant heterogeneity in national insurance systems and recognition rules, which often penalize countries with structured and inclusive reporting systems. Across the countries, musculoskeletal disorders, cancers, and respiratory diseases remain primary concerns, with annual incidence rates varying widely from 1.71 to 1,387 per 100,000 employees depending on the reporting mechanism used [5].

In Italy, occupational diseases are recorded through the national insurance system managed by the Italian National Institute for Insurance against Accidents at Work (INAIL). The system collects administrative and medico-legal information on occupational disease claims, including diagnoses, causal agents, economic sectors, and compensation outcomes. Over

the past decade, INAIL data have documented an increase in reported occupational disease claims, with musculoskeletal disorders accounting for the largest share [9]. Studies of Italian reporting pathways also show that missed recognition and delayed reporting remain important problems [10]. For this reason, INAIL claims data are best interpreted as data on registered and assessed claims, not as a direct estimate of occupational disease incidence in the whole workforce [11].

Beyond claim counts, impairment severity is an important dimension for occupational health surveillance. In the Italian system, the degree of impairment reflects permanent damage to a worker's psychophysical integrity and directly influences compensation and long-term support [12].

Understanding which factors are associated with higher impairment levels can support medico-legal evaluation, prevention planning, and resource allocation.

Previous Italian studies have mainly described occupational disease trends or specific disease groups [13, 14]. These analyses are valuable, but severity among compensated or assessed claims may depend on several factors at once, including demographic characteristics, diagnostic categories, causal agents, and occupational sectors. Traditional regression methods are useful for interpretation, but they may not capture non-linear relations among these variables.

Machine learning has been applied in occupational safety and health to predict injury, identify risk patterns, and support prevention research [15, 16]. Explainable and neural-network approaches have also been explored for musculoskeletal disorders and occupational disease prediction [17, 18]. In compensation systems, however, predictive models require careful interpretation because administrative variables may reflect recognition rules as well as clinical severity.

The present study addresses these gaps by combining descriptive epidemiology and predictive modelling in INAIL occupational disease claims from 2020 to 2024. We first describe the distribution of claims by demographic characteristics, economic sector, diagnostic category, and causal agent. We then identify factors associated with impairment severity through multivariable linear regression. Finally, we test whether embedding-based neural-network models can classify claims with recognized impairment and predict impairment severity as a proof of concept for administrative support.

Throughout the analysis, the unit of observation is the insurance claim. The findings therefore describe the profile and severity of claims handled within the INAIL system rather than the risk of developing an occupational disease in the Italian workforce.

Methods

STUDY DESIGN AND DATA SOURCE

This study is a retrospective observational analysis based on administrative data provided by INAIL, the public institution responsible for compulsory insurance

against occupational injuries and diseases in Italy. We used open-access data on occupational disease claims registered between January 1, 2020 and December 31, 2024. The INAIL occupational disease archive records claims submitted for medico-legal assessment and compensation. For each claim, the database may include worker demographics, country of birth, economic sector, ICD-10 diagnostic classification, suspected or recognized causal agent, asbestos-related status, scheduled disease status, compensated days, and the final degree of recognized impairment.

Because the database is built for insurance and compensation purposes, it does not include the total number of workers at risk in each sector and it does not capture unreported or unrecognized occupational diseases. It is therefore suitable for describing registered claims and their outcomes, but not for estimating disease incidence or relative risk in the general workforce.

The use of anonymized and publicly available data exempted the study from ethical approval requirements.

STUDY POPULATION

The initial dataset included 321,744 occupational disease claims recorded during the study period. Records with missing information on the degree of impairment were excluded, as this variable was the primary outcome of interest. After data preprocessing and cleaning, the final analytical sample consisted of 310,189 claims.

Each record corresponded to a single occupational disease claim and included information on the worker, the disease diagnosis, and the administrative outcome of the claim.

VARIABLES

The primary outcome was the degree of impairment, expressed as a percentage ranging from 0 to 100, representing the permanent damage to the worker's psychophysical integrity as recognized by INAIL [12]. This measure is used in the Italian compensation system to determine eligibility for benefits and the level of compensation. For analytical purposes, impairment was treated as a continuous variable in regression analyses. Independent variables were selected based on their epidemiological relevance and availability in the dataset and were grouped into four main domains:

1. Demographic and administrative characteristics, including sex, country of birth (Italy vs. abroad), classification as a scheduled (tabulated) occupational disease [19, 20], asbestos-related disease status, number of compensated days off work, and time to impairment definition ("days to definition" expressed as the number of days between claim submission and final impairment assessment).
2. Diagnostic categories, defined according to the International Classification of Diseases, 10th Revision (ICD-10).
3. Causal agents, categorized into macro-groups representing different types of occupational exposures, including chemical, physical, biological, ergonomic, psychological, and environmental factors.

4. Economic sector, classified into agriculture and fishing, industry, services, craftsmanship, public administration, and cases with unspecified sector.

Categorical variables were encoded using binary indicators for regression analyses.

Data preprocessing included consistency checks, exclusion of records with missing outcome data, and aggregation of diagnostic and causal variables into meaningful macro-categories to reduce dimensionality and improve interpretability. Continuous variables were examined for distributional properties, and descriptive statistics were used to identify outliers and data anomalies.

STATISTICAL ANALYSIS

Descriptive statistics were used to characterize the study population. Continuous variables were summarized using means, medians, standard deviations, and ranges, while categorical variables were described using absolute and relative frequencies. To identify determinants of impairment severity, a multivariable linear regression model was applied with degree of impairment as the dependent variable. Independent variables included demographic characteristics, diagnostic categories, causal agents, and economic sectors. Model performance was assessed using the coefficient of determination (R^2), and statistical significance was evaluated using p-values and 95% confidence intervals. This approach allowed estimation of the independent contribution of each predictor to impairment severity while controlling for potential confounding. All statistical analyses were conducted using Jamovi and Python, employing standard data science libraries including pandas, numpy, and scikit-learn. Alternative modeling strategies were considered; however, linear regression was preferred for clarity and comparability with previous epidemiological studies.

PREDICTIVE MODELING

As a secondary analysis, two predictive models were developed using neural networks with embedding layers: a classification model to distinguish occupational disease claims with recognized impairment ($>0\%$) from those without (0%), and a regression model to predict the specific degree of impairment among recognized cases. This dual formulation was chosen to support both case prioritization and severity estimation in occupational health surveillance.

Given the high cardinality of key categorical variables, particularly ICD-10 diagnostic codes and causal agents, a neural network architecture with embedding layers was implemented. This approach was selected to overcome the limitations of traditional encoding strategies, such as one-hot encoding, which would have substantially increased dimensionality and model complexity. Embedding layers allow each category to be represented as a dense, low-dimensional vector, enabling the model to capture latent similarities among categories and complex non-linear relationships.

High-cardinality categorical variables (ICD-10 codes and

causal agents) were encoded using label encoding and processed through embedding layers, while demographic and administrative variables were included as binary features. The feature set selected for model training specifically included: asbestos-related status, scheduled (tabulated) disease classification, sex, country of birth (Italy vs. abroad), ICD-10 diagnostic codes, causal agents, and economic sectors (unspecified, agriculture and fishing, industry, services, craftsmanship, and public administration). This variable subset was chosen to avoid data leakage, intentionally excluding features that are determined concurrently with or after the assessment of impairment, such as compensated days or the duration of the administrative process. Duplicate records were also removed prior to model training to prevent information leakage between training and validation datasets and to improve generalizability.

After preprocessing, approximately 250,000 observations were retained. For the regression task, training was restricted to cases with impairment $>0\%$. Both models used stratified sampling (90% training, 10% validation). Training was performed for 200 epochs using the Adam optimizer ($\eta=1\times10^{-3}$). The classification model used a weighted CrossEntropyLoss to account for class imbalance, while the regression model employed the Huber Loss to robustly handle outliers in impairment scores. Classification performance was evaluated using accuracy, precision, recall, and F1-score, while regression performance was assessed using R^2 , Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

Results

After data cleaning and exclusion of records with missing information on impairment, the final analytical sample consisted of 310,189 occupational disease claims registered between 2020 and 2024. Men accounted for 73.4% of claims and women for 26.6%. Most claimants were born in Italy (92.0%), while 8.0% were born abroad. More than half of claims had an unspecified economic sector. Among claims with a defined sector, industry accounted for approximately 15%, followed by craftsmanship (14%), services (11%), and agriculture and fishing (10%); public administration represented less than 1%.

Musculoskeletal disorders were the largest ICD-10 diagnostic group among claims. Other frequent groups included diseases of the ear and mastoid process, disorders of the nervous system, and diseases of the respiratory system. Neoplasms and hematological diseases were less frequent but were associated with greater impairment. The distribution of causal agents was dominated by ergonomic and physiological factors, followed by claims in which the causal agent was still undetermined. Material agents, industrial products, and physical agents were also common.

Chemical and biological agents accounted for smaller shares of claims but were relevant in specific occupational contexts. Asbestos-related diseases

Tab. I. Distribution of compensated days and recognized impairment severity among occupational disease claims.

Variable	Mean	Median	SD	Min	Max
Compensated days (>0)	70.1	55	52.6	1	342
Impairment severity (%)	2.2	0	5.9	0	100

Tab. II. Multivariable linear regression analysis of factors associated with recognized impairment severity among occupational disease claims. Dependent variable: degree of impairment (%); $R^2 = 0.551$; $n = 310,189$.

Variable	β	SE	95% CI	p
Demographic and administrative variables				
Days to definition (per day)	0.001	<0.001	0.001 to 0.002	<0.001
Asbestos-related disease (yes vs no)	13.053	0.167	12.726 to 13.379	<0.001
Sex (male vs female)	0.208	0.021	0.166 to 0.250	<0.001
Country of birth (Italy vs abroad)	-0.115	0.033	-0.179 to -0.050	<0.001
Scheduled disease (yes vs no)	-0.266	0.024	-0.313 to -0.219	<0.001
Diagnostic category (ICD-10; reference: absent)				
Infectious and parasitic diseases	2.823	1.373	0.132 to 5.513	0.040
Neoplasms	16.401	0.986	14.469 to 18.333	<0.001
Blood diseases	5.112	1.169	2.821 to 7.403	<0.001
Endocrine diseases	4.808	1.218	2.420 to 7.195	<0.001
Mental disorders	5.087	0.995	3.138 to 7.036	<0.001
Nervous system diseases	4.291	0.986	2.359 to 6.223	<0.001
Eye diseases	4.360	1.051	2.300 to 6.421	<0.001
Ear diseases	5.111	0.986	3.179 to 7.044	<0.001
Circulatory system diseases	4.991	1.001	3.029 to 6.952	<0.001
Respiratory system diseases	-5.852	0.987	-7.787 to -3.917	<0.001
Digestive system diseases	4.758	1.021	2.758 to 6.758	<0.001
Skin diseases	-0.520	0.993	-2.467 to 1.426	0.600
Musculoskeletal system diseases	5.280	0.985	3.349 to 7.211	<0.001
Genitourinary system diseases	5.086	0.985	3.155 to 7.017	<0.001
Symptoms / signs	4.372	5.061	-5.548 to 14.292	0.388
Causal agent (reference: absent)				
Organic chemical agents	4.734	1.001	2.771 to 6.697	<0.001
Inorganic chemical agents	10.503	1.046	8.454 to 12.552	<0.001
Biological agents	10.947	0.999	8.990 to 12.905	<0.001
Industrial materials and products	11.097	0.976	9.183 to 13.010	<0.001
Physical agents	-1.122	0.964	-3.012 to 0.767	0.244
Physiological factors	-1.693	0.964	-3.581 to 0.196	0.079
Psychological factors	-0.330	1.044	-2.377 to 1.717	0.752
Agent yet to be determined	-4.691	0.976	-6.603 to -2.778	<0.001
Not determined	-3.380	0.182	-3.737 to -3.023	<0.001
Economic sector (reference: unspecified)				
Agriculture and fishing	-1.546	0.088	-1.718 to -1.374	<0.001
Industry	-1.405	0.087	-1.575 to -1.235	<0.001
Services	-1.101	0.088	-1.274 to -0.928	<0.001
Craftsmanship	-1.379	0.087	-1.549 to -1.208	<0.001

β : unstandardized regression coefficient (percentage-point change in impairment severity).

represented approximately 2% of the total sample (5,913 claims). Overall, 27.1% of claims were classified as scheduled occupational diseases, while 72.9% were non-scheduled.

Summary statistics for the continuous outcome and compensation variables are shown in Table I. Compensated days were right-skewed, with a mean of 70.1 days among claims with at least one compensated absence. Impairment severity had a mean of 2.2% and a

median of 0%, confirming that most registered claims did not result in permanent recognized impairment. Values nevertheless ranged up to 100%, indicating a small group of highly severe claims.

Results of the multivariable linear regression analysis are reported in Table II. The model explained 55.1% of the variability in impairment severity. Among diagnostic categories, neoplasms showed the strongest association with impairment severity, with an average increase of

Tab. III. Global performance metrics of the predictive model (validation set).

Dataset	Accuracy	Precision	Recall	F1-Score
Validation	98.15%	98.21%	98.15%	98.15%

Class-specific performance metrics are reported in Table IV. The model showed particularly high recall for cases with recognized impairment, while maintaining excellent precision for cases without impairment.

16.4 percentage points compared with claims outside that category ($p < 0.001$). Positive associations were also observed for hematological, endocrine, neurological, ear, circulatory, musculoskeletal, genitourinary, and mental disorder categories. Among causal agents, asbestos-related status was associated with an increase of 13.1 percentage points. Inorganic chemical agents, biological agents, and industrial materials/products were each associated with increases greater than 10 percentage points. Male sex showed a small positive association, while being born in Italy and having a scheduled disease classification showed small negative associations. Economic sector indicators had modest negative coefficients after adjustment for diagnostic and causal factors, supporting the interpretation that claim severity was driven mainly by diagnosis and exposure category rather than sector alone.

Diagnostic category, causal agent, and economic sector variables are binary (1 = present/yes; reference = absent/no). Economic sector reference category: unspecified sector. CI: confidence interval.

The moderate explanatory power of the linear regression model suggested that non-linear methods could capture additional structure in the claim data. In the secondary predictive analysis, a neural network-based classification model was developed to distinguish claims with recognized impairment (> 0%) from those without impairment (0%). After removal of duplicate records, approximately 250,000 observations

were used, with 90% allocated to training and 10% to validation using stratified sampling.

The distribution analysis showed a relatively balanced dataset, with 114,385 claims without recognized impairment (44.52%) and 142,546 claims with recognized impairment (55.48%). To account for this mild class imbalance and to optimize performance across both outcome categories, a weighted loss function was adopted during training (weighted CrossEntropyLoss), assigning a weight of 1.12 to the minority class and 0.90 to the majority class.

Model training was conducted over 200 epochs using the Adam optimizer with a learning rate of 1×10^{-3} . The training process exhibited stable convergence without significant overfitting, monitored via validation metrics to ensure generalizability.

Overall model performance on the validation set was high (Tab. III), with accuracy exceeding 98% and similarly elevated values for precision, recall, and F1-score.

The confusion matrix (Fig. 1) confirms the model's strong ability to correctly identify cases resulting in recognized impairment, with a sensitivity of 99.99% and a very low number of false negatives. This characteristic is particularly relevant from both clinical and administrative perspectives, as it minimizes the risk of underestimating cases requiring compensation.

When compared with the multivariable linear regression

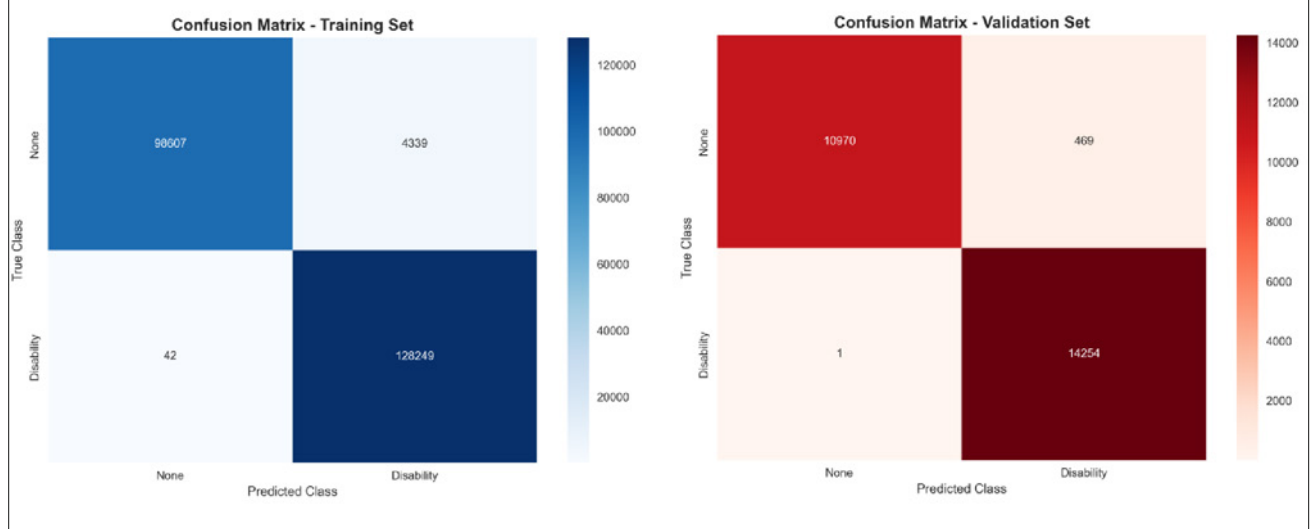
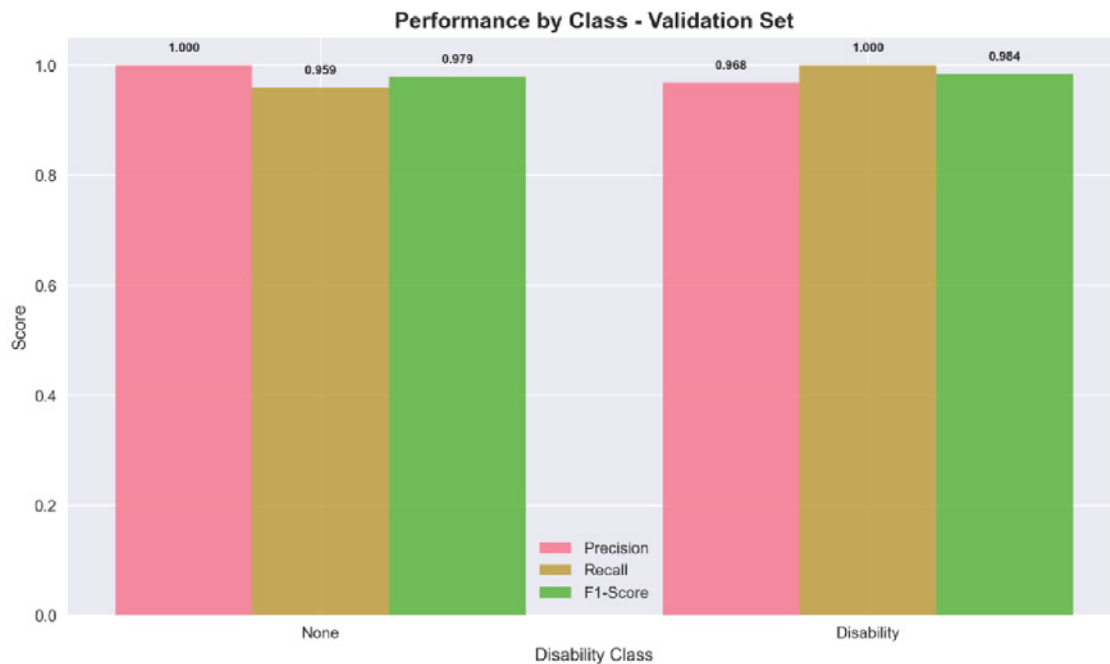
Fig. 1. Confusion matrices of the predictive model for training set (left) and validation set (right).

Fig. 2. Precision, recall, and F1-score for impairment classification on the validation set.**Tab. IV.** Class-specific performance of the predictive model (validation set).

Class	Precision	Recall	F1-Score
No impairment	99.98%	95.87%	97.88%
Recognized impairment	96.79%	99.99%	98.36%

Tab. V. Performance metrics of the regression model (validation set).

Metric	Value
MSE	30.37
RMSE	5.51%
MAE	2.31%
R^2	0.71

model described previously, the neural network with embedding layers demonstrated a superior capacity to capture complex, non-linear relationships among predictors. The model's classification performance is summarized in Figure 2. The embedding-based approach proved especially effective in handling high-cardinality categorical variables, overcoming limitations associated with traditional encoding techniques. Although the binary classification model does not predict the exact degree of impairment, it provides a more robust and operationally meaningful decision-support output, with direct implications for case management and occupational health surveillance.

The regression model, trained on cases with recognized impairment to predict the specific degree of severity, demonstrated strong predictive capability. On the validation set, the regression model similarly achieved robust performance (Tab. V), with a Coefficient of Determination (R^2) of 0.71. This represents a substantial improvement over the linear regression model ($R^2=0.55$) presented in Table II, highlighting the neural network's ability to capture non-linear interactions between risk factors. The Mean Absolute Error (MAE) of 2.31% indicates that, on average, the model's prediction deviates from the actual impairment degree by less than 2.31 percentage points.

Discussion

This study describes occupational disease compensation claims registered by INAIL in Italy between 2020 and 2024. By combining descriptive analysis, multivariable regression, and predictive modelling, it identifies which claim characteristics were most strongly associated with recognized impairment. The findings should be read as patterns within the insurance claims system, not as estimates of disease risk in the workforce.

Consistent with previous national and international studies [9, 14], musculoskeletal disorders represented the largest proportion of occupational disease claims. This predominance reflects the persistent exposure of workers to ergonomic risk factors such as repetitive movements, manual handling of loads, and prolonged awkward postures. Despite their high frequency, musculoskeletal disorders were generally associated with lower levels of impairment compared with other diagnostic groups, suggesting that their overall impact on permanent disability is substantial mainly due to volume rather than individual severity.

In contrast, neoplasms and asbestos-related diseases, although accounting for a smaller proportion of claims, were associated with markedly higher impairment

levels. The strong association observed between asbestos exposure and impairment severity confirms the long-term health consequences of past occupational exposures and underscores the continued relevance of asbestos-related surveillance, even decades after regulatory bans. Similarly, the elevated impairment associated with chemical and biological agents highlights the need for targeted preventive measures in high-risk occupational settings, including healthcare and industrial environments.

THE PARADOX OF FREQUENCY AND SEVERITY

A striking finding of this analysis is the inverse relationship between claim frequency and clinical severity. Musculoskeletal disorders (MSDs) dominate the volume of claims (>65%) yet show low average impairment. This phenomenon can be partially explained by the “Healthy Worker Survivor Effect”: workers with severe MSDs may leave the workforce early or transition to disability pensions (INPS) rather than filing continuous INAIL claims for permanent impairment. Conversely, the high severity of neoplasms (coefficient +16.4%) reflects the aggressive nature of occupational cancers (e.g., mesothelioma, sinonasal carcinoma), which often present at advanced stages due to long latency periods and lack of early symptoms. The high impairment scores in this group are unfortunately often a proxy for poor prognosis and reduced life expectancy, necessitating expedited compensation pathways.

IMPLICATIONS FOR SOCIAL DIALOGUE AND LEGISLATION

The integration of predictive insights into the “social dialogue” is crucial for modernizing occupational health governance in Italy. The granular risk profiling provided by this study can inform negotiations between trade unions and employer associations, facilitating the development of targeted prevention measures in collective bargaining agreements. This approach aligns with the principles of legislative Decree 81/2008 [21], which emphasizes participatory risk assessment. By identifying high-severity drivers, such as specific material agents in the industrial sector, stakeholders can move beyond generic safety protocols to implement evidence-based interventions. Furthermore, the ability to predict impairment probability can assist outcomes-based policy planning, allowing regional health authorities and INAIL to allocate resources where the burden of permanent disability is highest.

PREDICTIVE MODELING AS A SUPPORT TOOL

As a secondary objective, this study explored the feasibility of predictive modeling to support occupational disease surveillance. The very high predictive performance observed is likely influenced not only by clinical severity patterns, but also by structured medico-legal and administrative rules embedded in the INAIL compensation process. In this context, the model may partially learn consistent decision-making patterns related to impairment recognition, rather than

purely clinical features. For this reason, the predictive analysis should be interpreted as a proof of concept for administrative support rather than as a diagnostic or clinical prediction tool. Specifically, the model could serve as a “triage engine” to flag high-probability impairment claims for expert review, thereby optimizing resource allocation in the compensation process.

The integration of traditional epidemiological methods with predictive modeling represents a promising approach for occupational safety and health research. While regression models offer interpretability and causal insight, machine learning techniques, specifically neural networks with embedding layers, capture complex, non-linear relationships that are difficult to model using conventional approaches. This is evidenced by the superior performance of the neural network regression model ($R^2=0.72$) compared to the linear baseline ($R^2=0.55$), suggesting that the determinants of impairment interact in non-additive ways. Their combined use may enhance both understanding and practical application, provided that transparency and validation requirements are adequately addressed.

LIMITATIONS

The main limitation is the claim-based design. The database does not provide workforce denominators by sector and region, so incidence rates and relative risks could not be calculated. The results describe registered and assessed claims, not the risk of disease in the Italian workforce.

Underreporting and delayed recognition of occupational diseases may also affect the composition of the dataset. The data were collected for compensation rather than research. Diagnoses, exposures, and economic sectors may therefore be incomplete or misclassified, and more than half of records lacked a specific economic sector. The outcome, degree of impairment, is a medico-legal measure rather than a pure clinical severity scale, and the distribution was zero-inflated because many claims had 0% recognized impairment. Finally, although duplicates were removed and leakage-prone variables were excluded from predictive modelling, external validation on independent data is needed before any operational use.

Conclusion

This study demonstrates the value of integrating national insurance administrative data with advanced analytical techniques to enhance occupational health surveillance in Italy. The findings confirm that while musculoskeletal disorders dominate the volume of claims, the highest burden of permanent impairment remains linked to neoplasms and asbestos-related diseases, highlighting the need for continued vigilance and targeted prevention for high-risk exposures.

The successful implementation of a neural network-based model to predict recognized impairment represents a significant step towards more proactive and data-driven

decision-making in occupational medicine. Despite limitations inherent in administrative datasets, such as underreporting and lack of detailed clinical variables, this integrated approach provides a robust framework for identifying high-risk cases and optimizing resource allocation within compensation systems.

Future developments should focus on incorporating explainable AI techniques to bridge the gap between predictive performance and clinical interpretability, ultimately supporting practitioners and policymakers in developing more effective strategies for worker protection.

Acknowledgements

The authors wish to thank INAIL for making occupational disease data publicly available through their open data portal.

Conflicts of interest

The authors declare no conflicts of interest.

Authors' contributions

MP conceived the study, performed the literature review, interpreted the results, and drafted the manuscript. AP designed and implemented the data analysis pipeline, developed the predictive models, and contributed to manuscript writing. Both authors read and approved the final manuscript.

References

- [1] Takala J, Härmäläinen P, Sauni R, Nygård CH, Gagliardi D, Neupane S. Global-, regional- and country-level estimates of the work-related burden of diseases and accidents in 2019. *Scand J Work Env Hea* 2024;50:73-82. <https://doi.org/10.5271/sjweh.4132>.
- [2] Härmäläinen P, Takala J, Kiat TB. Global estimates of occupational accidents and work-related illnesses 2017. Technical report, Workplace Safety and Health Institute, Singapore. Available at: <https://www.icohweb.org/site/images/news/pdf/Report%20Global%20Estimates%20of%20Occupational%20Accidents%20and%20Work-related%20Illnesses%202017%20rev1.PDF>. Accessed on: 27/12/2025.
- [3] World Health Organization and International Labour Organization. Who/ilo joint estimates of the work-related burden of disease and injury, 2000–2016: Global monitoring report. Technical report, World Health Organization and International Labour Organization, Geneva, 2021. Available at: <https://www.who.int/publications/i/item/9789240034945>. Accessed on: 09/01/2026.
- [4] Azaroff LS, Levenstein C, Wegman DH. Occupational injury and illness surveillance: Conceptual filters explain underreporting. *Am J Public Health* 2002;92:1421-9. <https://doi.org/10.2105/AJPH.92.9.1421>
- [5] Khoe LC, Fitriana Saldi SR, Isbayuputra M, Mansyur M, Wiseman V, Asante A. Global reporting and underreporting of occupational diseases: A systematic review. *PloS one* 2026;21:e0345318. <https://doi.org/10.1371/journal.pone.0345318>.
- [6] Nagy K, Kudász F. Introduction to occupational diseases. OS-Hwiki, European Agency for Safety and Health at Work, 2022. Published 08 April 2014, last updated 09 September 2022. Available at: <https://oshwiki.osha.europa.eu/en/themes/introduction-occupational-diseases>. Accessed on: 15/01/2026.
- [7] European Agency for Safety and Health at Work. Work-related diseases and occupational diseases: The issue. Technical report, Publications Office of the European Union, Luxembourg, 2017. Available at: <https://osha.europa.eu/en/publications/work-related-diseases-and-occupational-diseases-issue>. Accessed on: 22/01/2026.
- [8] Eurostat. European occupational diseases statistics (EODS). Experimental Statistics, European Commission, 2025. Available at: <https://ec.europa.eu/eurostat/web/experimental-statistics/european-occupational-diseases-statistics>. Accessed on: 12/01/2026.
- [9] Larese Filon F, Spadola O, Colosio C, Van Der Molen H. Trends in Occupational Diseases in Italian Industry and Services From 2006 To 2019. *Med Lav* 2023;114:e2023035. <https://doi.org/10.23749/mdl.v114i4.14637>.
- [10] European Agency for Safety and Health at Work. Methodologies to identify work-related diseases – Review of sentinel and alert approaches, Publications Office, 2017. Available at: <https://data.europa.eu/doi/10.2802/053155>. Accessed on: 15/01/2026.
- [11] Lecca LI, Pili S, Lai M, Murru A, Campo G, Pizzuti A, Mattioli S, Campagna M. Diagnosing and Reporting of Occupational Diseases: An Assessment Study of Reports from an Italian Workplace Safety Prevention Program Service. *Med Lav* 2025;116:16609. <https://doi.org/10.23749/mdl.v116i1.16609>.
- [12] Decreto legislativo. Disposizioni in materia di assicurazione contro gli infortuni sul lavoro e le malattie professionali (d.lgs. 23 febbraio 2000, n. 38), 2000. *Gazzetta Ufficiale* n. 60, 13 marzo 2000. Available at: <https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2000;38-art12>. Accessed on: 21/12/2025.
- [13] Ferrante P. Respiratory Diseases with High Occupational Fraction in Italy: Results from the Italian Hospital Discharge Registry (2010-2021). *Healthcare (Basel)* 2024;12:2565. <https://doi.org/10.3390/healthcare12242565>.
- [14] van der Molen HF, Marsili C, Vitali A, Colosio C. Trends in occupational diseases in the Italian agricultural sector, 2004-2017. *Occup Environ Med* 2020;77:340-3. <https://doi.org/10.1136/oemed-2019-106168>.
- [15] Vivian GA, Bauder RA, Khoshgoftaar TM. A comprehensive survey on machine learning for workplace injury analysis: risk prediction, return to work strategies, and demographic insights. *J Big Data* 2025;12:167. <https://doi.org/10.1186/s40537-025-01229-z>.
- [16] Armenteros-Cosme P, Arias-González M, Alonso-Rollán S, Márquez-Sánchez S, Carrera A. Advancements in Artificial Intelligence and Machine Learning for Occupational Risk Prevention: A Systematic Review on Predictive Risk Modeling and Prevention Strategies. *Sensors* 2025;25:5419. <https://doi.org/10.3390/s25175419>.
- [17] Luo N, Xu X, Jiang B, Zhang Z, Huang J, Zhang X, Tan Q, Wang X, Bai S, Liu S, Pan Y, Tang, C, Zhu P. Explainable machine learning framework to predict the risk of work-related neck and shoulder musculoskeletal disorders among healthcare professionals. *Front Public Health* 2024;12:1414209. <https://doi.org/10.3389/fpubh.2024.1414209>.
- [18] Montanari P. Occupational Diseases Risk Prediction by Neural Networks. *Health* 2025;17:579-93. <https://doi.org/10.4236/health.2025.175037>.
- [19] Decreto interministeriale. Nuove tabelle delle malattie professionali nell'industria e nell'agricoltura (9 aprile 2008), 2008. *Gazzetta Ufficiale* n. 169, 21 luglio 2008. Available at: https://www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2008-07-21&atto.codiceRedazionale=08A05160&elenco30giorni=false. Accessed on: 12/12/2025.

- [20] Decreto ministeriale. Aggiornamento delle tabelle delle malattie professionali (10 ottobre 2023), 2023. Gazzetta Ufficiale n. 269, 18 novembre 2023. Available at: <https://www.gazzettaufficiale.it/eli/id/2023/11/18/23A06309/sg>. Accessed on: 21/01/2026.
- [21] Decreto legislativo. Attuazione dell'articolo 1 della legge 3 agosto 2007, n. 123, in materia di tutela della salute e della sicurezza nei luoghi di lavoro (d.lgs. 9 aprile 2008, n. 81), 2008. Gazzetta Ufficiale n. 101, 30 aprile 2008. Available at: <https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2008-04-09;81!vig=>. Accessed on: 24/03/2026.

Received on February 14, 2026. Accepted on July 24, 2026.

Correspondence: Martina Padovan, Preventive Medicine, Tuscany North-West Health Local Unit, Italy. E-mail: padovan.martina@gmail.com

How to cite this article: Padovan M, Palla A. Occupational disease compensation claims in Italy: an epidemiological analysis of national insurance data with predictive insights. J Prev Med Hyg 2026;67:E409-E417. <https://doi.org/10.82082/2421-4248/jpmh2026.67.3.3873>

© Copyright by Pacini Editore Srl, Pisa, Italy

This is an open access article distributed in accordance with the CC-BY-NC-ND (Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International) license. The article can be used by giving appropriate credit and mentioning the license, but only for non-commercial purposes and only in the original version. For further information: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>