safety

MDPI

Article

Can Occupational Safety Practitioners Trust AI-Ranked Risk Factors? A Synthetic Study of Predictive Performance and Explanation Fidelity

Raúl Aguilar-Elena [1], Ana Delgado-García [2], Rubén Tino-Ramos [1], Alejandro Guillén-Riquelme [1,*] and Juan José Agún-González [1]

[1] School of Social and Legal Sciences, Universidad Internacional de Valencia (VIU), 46002 Valencia, Spain; raguilar@universidadviu.com (R.A.-E.); ruben.tino@universidadviu.com (R.T.-R.); jjagun@universidadviu.com (J.J.A.-G.)
[2] BISITE Research Group, Universidad de Salamanca, 37008 Salamanca, Spain; a.delgado@usal.es
* Correspondence: aguillenr@universidadviu.com

Abstract

Artificial intelligence tools are increasingly used in occupational safety to rank factors associated with incidents and to support decisions on training, supervision, inspection, and preventive resource allocation. We evaluated whether predictive performance and explanation fidelity remain aligned when events are rare, and predictors are correlated. A fully synthetic occupational system was generated with Safety Data Sheet competence, six Bloom's Taxonomy dimensions, behavioral and organizational variables, and a simulated incident outcome. Random forests and ridge models were assessed for nested cross-validation discrimination and calibration; random forests were additionally evaluated using ablation, held-out marginal and conditional permutation feature importance, and matched oracle importance derived from the known data-generating mechanism. A 29-condition Monte Carlo experiment (58,000 random-forest fits) examined sample size, prevalence, effect magnitude, and predictor architecture. Under the baseline condition, mean AUC was 0.689, whereas oracle–model rank agreement was weak (Spearman $\rho = 0.048$) and oracle top-1 recovery was 20.7%. A high-signal implementation check produced $\rho = 0.964$, with 83.4% top-1 and 98.5% top-2 recovery, while a more intensive rare-event pipeline left top-1 recovery essentially unchanged at 20.3%. Conditional permutation importance, benchmarked against a matched conditional oracle, reduced spurious first-ranking of true-null controls (26.9% to 22.9%, $p < 0.001$) but did not improve recovery of the oracle top variable (20.4% versus 20.7%). AI-derived factor rankings should therefore be validated separately from predictive performance before they are used to set preventive priorities within conventional occupational risk-assessment processes. The evidence presented here concerns random forests with marginal and conditional permutation importance under fully synthetic rare-event conditions, and does not generalize to AI explanation methods as a class.

Keywords: occupational safety; occupational health; explainable artificial intelligence; XAI; permutation feature importance; synthetic data; Safety Data Sheets; Bloom taxonomy; rare events; risk prioritization

Check for updates

Academic Editor: Raphael Grzebieta

Received: 16 September 2026
Revised: 22 October 2026
Accepted: 4 November 2026
Published: 12 November 2026

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

1. Introduction

Machine-learning models are increasingly used to identify patterns associated with occupational incidents, unsafe conditions, and human-error precursors. Recent reviews

show that occupational AI has moved beyond accident classification towards predictive risk modelling, targeted prevention, and decision support, while repeatedly identifying interpretability, data quality, class imbalance, and translation into practice as unresolved challenges [1,2]. Contemporary studies integrate machine learning with inspection data, explainable modelling, and automated ergonomic assessment to support proactive decision-making [3–5]. The practical consequence is that model outputs are no longer confined to academic prediction exercises: internally developed dashboards and commercial systems now present supervisors, safety professionals, and managers with ranked lists of factors labelled "most important", "highest risk", or "main drivers". A preventive professional with limited time and resources must decide where to intervene first, and such a ranking is an obvious candidate for making that decision.

The methodological difficulty is that a model's ability to distinguish higher-risk from lower-risk observations does not imply that its explanation method has correctly recovered the structure responsible for those predictions. A model can achieve a respectable area under the receiver operating characteristic curve (AUC-ROC) and still produce variable-importance estimates that are unstable, distorted by sparse events, or sensitive to correlated predictors. Existing occupational-safety reviews have largely evaluated which models are used and how accurately they predict [1,2]; much less attention has been paid to the separate question that arises after prediction, namely whether the factor ranking supplied to the practitioner faithfully recovers the predictive-importance structure it claims to describe. This matters particularly in safety-critical settings, where interpretability is a prerequisite for responsible use of AI rather than an optional presentation layer [6,7], and recent work in *Safety* has likewise emphasized validation, explainability, robustness, and human oversight when AI supports occupational-risk prevention [8]. Predictive performance and explanation fidelity therefore need to be evaluated as separate properties.

One distinction should be made at the outset, because it governs how any ranking in this paper may be read. Variable importance is a property of the fitted model and the chosen performance criterion, not an intrinsic causal property of the variable itself [9]. A variable may rank highly because it carries predictive information, because it proxies for something else, or because of the dependence structure among predictors. None of these implies that intervening on that variable would reduce incidents. Predictive importance and intervention importance are different quantities, and a ranking that is accurate in the first sense can still be a poor guide to preventive action. Throughout this study, "importance" means predictive importance only.

Permutation feature importance (PFI) is attractive in applied settings because it is model-agnostic and easy to communicate: a feature is perturbed, and the resulting deterioration in predictive performance is measured [10,11]. Its value nonetheless depends on the fitted model, the loss function, and the joint distribution of the predictors. When predictors are correlated, unrestricted permutation generates feature combinations that are uncommon or implausible in the original data, forcing extrapolation and redistributing apparent importance [12,13]. Permutation-based null distributions have been proposed to separate informative from non-informative variables [14], and recent guidance stresses that different importance measures answer different inferential questions [15]. These are known limitations in the methodological literature; what is not established is how severe they become under the event-prevalence characteristic of real occupational incident data.

Assessing an importance method also requires defining the target of recovery correctly. The raw coefficients of a data-generating equation are not equivalent to predictive permutation importance, because the latter additionally depends on predictor variance, dependence, interactions, and the fitted prediction rule. A valid benchmark must therefore compare an estimated importance with an oracle version of the same quantity, obtained under the

same intervention, rather than with the magnitude of programmed coefficients. The formal definitions are given in Section 2.10.

Synthetic data make such a benchmark possible, because the outcome-generating mechanism is fully known. Methodological work emphasizes that synthetic data are purpose-built data generated by a mathematical model or algorithm, and are especially useful when controlled validation or otherwise unavailable ground-truth is required, provided their intended use and fidelity are made explicit [16–19]. The data used here are entirely computer-generated. No human participants, workers, or identifiable records were involved, and no empirical claim about any real workforce is made. Synthetic data serve a single purpose in this study: as a controlled validation instrument in which structural fidelity can be checked directly, true-null variables can be inserted deliberately, and oracle predictive loss can be recomputed after controlled perturbations.

The substantive setting is hazard communication through Safety Data Sheets (SDS), which are central instruments under the Globally Harmonized System and national frameworks derived from it [20–22]. Access to an SDS does not imply that a worker can retrieve, understand, interpret, and apply its contents under operational conditions, and reviews report recurring limitations in SDS accuracy, readability, and workplace usability [23,24]. The problem is acute in high-hazard industries such as oil and gas, where chemical exposure, non-routine tasks, time pressure, and organizational conditions coexist [25], and where competency-based training has long been favored over one-off information delivery [26–28]. Bloom's Taxonomy supplies a structured representation of this competency, distinguishing progressively more complex cognitive operations—remember, understand, apply, analyze, evaluate, and create/adapt [29,30]—a distinction that is operationally meaningful, since recalling a pictogram is not equivalent to selecting controls during an unexpected transfer operation. Bloom-based approaches have been applied in other training contexts [31,32], and higher-order training can outperform passive information delivery [33]. Here, these correlated cognitive dimensions provide a realistic and theoretically structured test case rather than a hypothesis about real training effects.

This study asks whether a practitioner-facing ranking of "important risk factors" remains trustworthy when incident outcomes are rare, and candidate predictors are correlated. Four objectives follow. First, we compare a unitary SDS-competency representation with a six-dimensional Bloom representation in the same synthetic workers and under the same realized outcome. Second, we examine the stability of held-out permutation importance at individual-variable and preventive-domain level, including genuine null controls with no pathway to the outcome. Third, we use an oracle benchmark under known ground-truth to quantify how faithfully fitted rankings recover the embedded predictive structure, applying this to both marginal and conditional permutation importance, each against its own matched oracle. Fourth, we translate the findings into a practical appraisal instrument for OSH professionals. Because the amount of outcome information constrains what can be estimated from a dataset [34], the analysis is organized around both event count and event prevalence as determinants of recovery. The study does not estimate causal effects of competency in real workers, and its conclusions are bounded by the estimators and synthetic conditions examined.

The novelty of this work lies in the integration rather than in any single component. Oracle-based benchmarking of importance estimators is established in the methodological literature, and the limitations of unrestricted permutation under predictor dependence are well documented [12,13,15]. What has not previously been reported, to our knowledge, is the systematic quantification of explanation-ranking recovery under occupational rare-event conditions: using a data-generating mechanism built around a substantive occupational-safety use case, with a Monte Carlo design large enough to estimate recovery probabilities

with useful precision, genuine null controls embedded by construction, two importance estimators each benchmarked against its own matched oracle, and an explicit translation of the resulting uncertainty into preventive decision support.

The contribution is therefore applied rather than algorithmic. It establishes how much confidence an occupational-safety professional can place in an AI-generated factor ranking under event prevalence typical of real incident data, and it supplies an appraisal instrument derived from that evidence rather than asserted alongside it.

2. Methods

2.1. Study Design

A computational simulation study was conducted using a single shared synthetic population of $n = 1800$ workers. The design generated two feature representations from the same underlying worker system. Dataset v1 represented SDS competency with one noisy unitary score (*sds_comprehension_score*). Dataset v2 represented competency through six correlated Bloom dimensions: *bloom_remember*, *bloom_understand*, *bloom_apply*, *bloom_analyze*, *bloom_evaluate*, and *bloom_create_adapt*.

The two datasets shared worker identifiers, context and organizational variables, four behavioral mediator variables, and the same Bernoulli-sampled incident outcome. Consequently, any difference in predictive performance between v1 and v2 reflects the representation available to the fitted model rather than a different population or different realized outcome.

The study used exclusively computer-generated data. No human participants or human records were involved.

Table 1 summarizes what the two representations share and where they differ.

Table 1. The two competency representations evaluated. Both were generated from one shared synthetic population and one realized outcome, so any difference in results is attributable to the competency representation alone.

	Dataset v1 (Unitary)	**Dataset v2 (Bloom-Expanded)**
Competency representation	One noisy unitary score, *sds_comprehension_score*	Six correlated Bloom dimensions: *remember, understand, apply, analyze, evaluate, create/adapt*
Underlying latent construct	Identical; generated once per worker	Identical; generated once per worker
Context and organizational variables	Shared	Shared (identical values)
Behavioral mediators	Four, shared numerically	Four, shared numerically
True-null controls	Two, no pathway to the outcome	Two, identical values
Outcome definition	Same Bernoulli draw from the same latent probability	Same Bernoulli draw from the same latent probability
Direct Bloom terms in the outcome equation	Present in the mechanism but not observable in this representation	Present and observable

The outcome equation contains explicit Bloom terms, so the mechanism is deliberately favorable to v2. A failure of v2 to outperform v1 therefore cannot be attributed to a mechanism that excluded Bloom information.

Figure 1 summarizes the complete study workflow, from generation of the synthetic occupational system through to the separate evaluation of predictive discrimination and explanation fidelity.

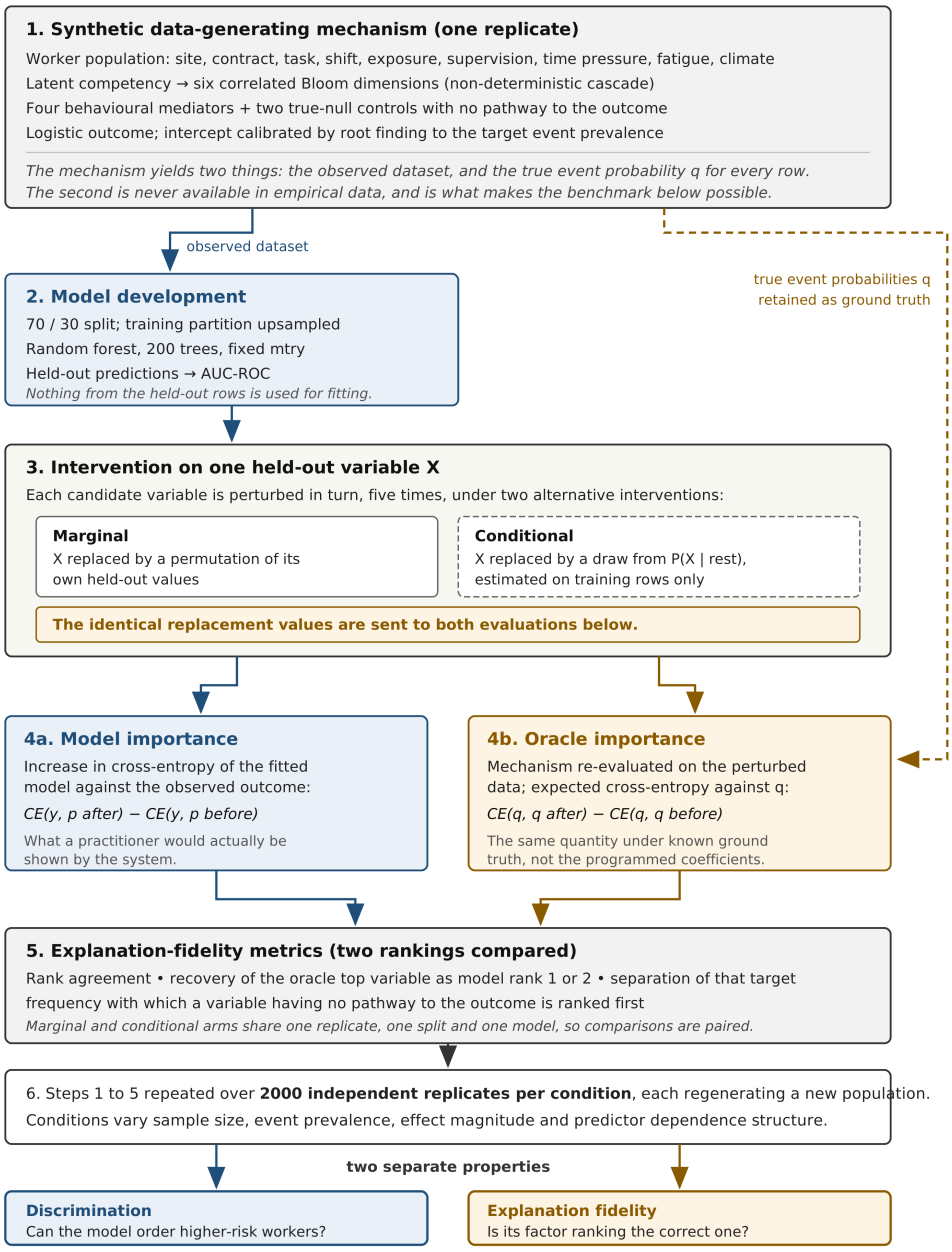


Figure 1. Study workflow. A synthetic occupational system generates both the observed dataset and, uniquely, the true event probability for every worker. A random forest is fitted to the training partition and evaluated on held-out observations. Each candidate variable is then perturbed under two alternative interventions, marginal permutation and conditional resampling, and the identical replacement values are supplied both to the fitted model and to the data-generating mechanism. This yields an estimated importance ranking and an oracle ranking of the same quantity, which are compared directly. The procedure is repeated across independent replicates and conditions. Discrimination and explanation fidelity are thereby estimated as separate properties of the same pipeline.

2.2. Synthetic Occupational System

The synthetic population included worksite, contractor type, job family, task type, shift, education, employment type, experience, language–SDS alignment, SDS training hours, chemical exposure level, procedure availability, permit-to-work status, safety climate, supervision, time pressure, fatigue, prior minor incidents, and non-routine-task status. Variables were generated on fixed, interpretable scales rather than being min–max rescaled within each generated sample.

A latent competency construct was generated once per worker. The v1 unitary score was a noisy measure of this latent construct. The v2 Bloom variables were generated as a non-deterministic cascade in which each dimension depended partly on latent competency, partly on the preceding cognitive dimension, and, where conceptually appropriate, on contextual modifiers such as supervision, permit-to-work availability, task non-routineness, experience, fatigue, safety climate, and language alignment. Independent Gaussian noise was added at each stage.

Four intermediate behavioral variables—PPE compliance, procedure adherence, hazard detection, and decision quality—were generated from the same latent competency construct and organizational/contextual factors. These variables were shared numerically between v1 and v2.

2.3. Outcome Data-Generating Mechanism

The binary outcome represented whether a simulated occupational incident occurred. Its latent probability was generated with a logistic model incorporating chemical exposure, higher-risk task categories, non-routine work, time pressure, fatigue, prior minor incidents, supervision, safety climate, the four behavioral mediators, three operational interaction terms, and explicit direct coefficients for the six Bloom dimensions.

The direct Bloom coefficients were fixed as follows:

$$\beta_{\text{remember}} = 0, \quad \beta_{\text{understand}} = -0.010, \quad \beta_{\text{apply}} = -0.012, \quad \beta_{\text{analyze}} = 0,$$

$$\beta_{\text{evaluate}} = -0.014, \quad \beta_{\text{create/adapt}} = -0.006.$$

Thus, *remember* and *analyze* had zero direct coefficients but were not true-null predictors because, as correlated members of the Bloom system, they could act as proxies for other predictive information. To provide genuine negative controls, two additional independent variables (*xai_negctrl_1* and *xai_negctrl_2*) were generated with no direct or indirect pathway to the outcome.

The intercept of the primary DGM was calibrated by numerical root finding so that the *expected* event prevalence was 3.6%. The Bernoulli draw was then allowed to vary naturally; the realized prevalence was not forced to equal the target. The final primary dataset contained 62 incidents among 1800 workers (3.44%).

2.4. Twin Representation and Falsification Scenarios

Datasets v1 and v2 were evaluated as alternative observed representations of the same underlying system. Importantly, the DGM was intentionally favorable to the Bloom-expanded representation because the outcome equation explicitly contained Bloom-dimension terms, whereas v1 contained only the unitary proxy. Failure of v2 to outperform v1 would therefore not be attributable to a DGM that excluded Bloom information.

Two direct-effect falsification scenarios were additionally generated. In the half-effect scenario, all direct Bloom coefficients were multiplied by 0.5; in the null-effect scenario, they were multiplied by 0. For the primary falsification comparison, the intercept was recalibrated in each scenario to maintain the same 3.6% expected prevalence. Separate fixed-intercept scenarios were retained as sensitivity analyses because they allow prevalence to change as a downstream consequence of altering the protective Bloom effect.

2.5. Predictive Modelling and Nested Cross-Validation

Two model classes were compared: ridge-penalized logistic regression and random forest [35,36]. Discrimination analyses used class upsampling within the training data to address the severe class imbalance. Hyperparameters were tuned only inside training partitions.

Primary discrimination was estimated using repeated nested cross-validation with five outer folds, three repeats, and five-fold inner cross-validation for tuning. Each worker therefore received one out-of-fold prediction per repeat. The three out-of-fold predictions were averaged at worker level before calculating the final AUC-ROC and precision–recall AUC (PR-AUC), avoiding pseudo-replication of workers across repeats.

AUC confidence intervals were generated from 2000 bootstrap resamples at worker level. Because v1 and v2 contained the same workers and outcomes, differences in AUC were evaluated using paired bootstrap resampling of the same worker indices. The primary contrast was

$$\Delta AUC = AUC_{v2} - AUC_{v1}.$$

2.6. Calibration

Class upsampling changes the outcome prevalence observed by the fitting algorithm and can distort absolute risk probabilities if probabilities are interpreted directly. Therefore, calibration was re-estimated in a separate nested cross-validation pipeline *without* class upsampling. Brier score, prevalence-only reference Brier score, Brier skill score, calibration intercept, and calibration slope were calculated from worker-level out-of-fold predictions.

This separation allows discrimination and probability calibration to be evaluated without attributing a prevalence-shift artefact to the model itself.

2.7. Ablation Analysis

The random-forest analysis evaluated six configurations: full v2; v2 without Bloom dimensions; v2 without Bloom dimensions and without the four behavioral variables; v2 without behavioral variables; Bloom dimensions only; and behavioral variables only. The contrast between full v2 and the configuration without Bloom plus behavioral descendants was evaluated using paired bootstrap inference on worker-level predictions.

This comparison is interpreted as incremental predictive contribution of the observed competency-related feature block. It is not interpreted as a causal decomposition of the total effect of competency.

2.8. Held-Out Permutation Importance

Permutation importance was evaluated on data not used to fit the explained model. Random forests were trained within outer training partitions and importance was calculated on the corresponding held-out test partitions using the increase in cross-entropy loss.

Individual variables were permuted one at a time. Domain importance was calculated by jointly permuting all variables assigned to the same domain rather than summing separate individual importances, following the logic of grouped permutation feature importance [37]. Domains were intermediate behaviors, Bloom competency, contextual factors, organizational conditions, training/SDS factors, and true-null controls.

2.9. Structural-Fidelity Checks

Four associations were specified before analysis to verify that the generated system behaved in the intended direction: fatigue with latent incident probability; Bloom *apply* with latent incident probability; supervision with procedure adherence; and language match with Bloom *understand*. Pearson correlation was used for continuous pairs and Spearman correlation for the ordinal language variable. These checks assess whether the DGM behaves as programmed; they do not test whether an explainability pipeline recovers that structure.

2.10. Oracle Permutation-Importance Benchmark

Permutation importance is not a causal-effect estimator. Its value depends on the fitted model, the loss function, the distribution of the predictors, correlations among them, sample size, and the amount of information carried by the outcome. For the same reasons, the raw coefficients of the data-generating equation are not equivalent to predictive permutation importance: coefficient magnitude, predictor variance, dependence structure, interactions, and the fitted prediction rule all affect how much predictive loss changes after a variable is perturbed. A valid benchmark must therefore compare an estimated importance with an oracle version of the same quantity, obtained under the same intervention, rather than with the absolute value of a programmed coefficient.

The central signal-recovery benchmark used an oracle version of permutation importance. Within each Monte Carlo replicate, the DGM itself provided the true event probability q_i for every held-out observation. For each variable X_j, the same held-out permutation used for model–PFI was applied to X_j, and the DGM probability was recomputed after permutation. Oracle importance was defined as the increase in expected cross-entropy:

$$I_j^{oracle} = CE(q, p^{perm(j)}) - CE(q, q),$$

where

$$CE(q, p) = -\frac{1}{n}\sum_{i=1}^{n}[q_i \log(p_i) + (1 - q_i)\log(1 - p_i)].$$

This definition compares model–PFI with the same predictive quantity under known ground-truth. Correlations with the absolute programmed coefficients were retained only as secondary diagnostics because coefficient magnitude and permutation importance are not equivalent estimands.

Recovery was summarized using mean Spearman ρ and Kendall τ between oracle and model Bloom importances; the rank assigned by the model to the oracle top variable; the probability that the oracle top variable was recovered as model top-1 or top-2; the probability that a zero-direct-coefficient Bloom dimension ranked first; and the probability that one of the two true-null controls ranked first among the six Bloom variables plus the two null controls.

2.11. Monte Carlo Sensitivity Analysis

The final Monte Carlo design used 29 conditions and 2000 independent replicates per condition (58,000 random-forest fits). Each replicate generated a new synthetic population and outcome, split observations randomly into 70% training and 30% held-out testing, upsampled the training data only, fitted a 200-tree random forest with fixed reasonable *mtry*, and calculated model and oracle-PFI using five repeated held-out permutations per variable.

Sixteen prevalence-standardized conditions used the correlated Bloom cascade and varied one factor at a time around the baseline ($n = 1800$, expected prevalence 3.6%, noise multiplier 1, correlation/loading multiplier 1, direct-effect multiplier 1): sample size (600, 1200, 3600), event prevalence (2%, 6%, 10%), noise (0.7, 1.4), cascade loading/correlation (0.7, 1.3), and direct-effect magnitude (0, 0.25, 0.5, 0.75, 1.5). Expected prevalence was recalibrated to the target in each of these conditions.

A second, independent-predictor comparator architecture removed the shared latent/cascade dependence among the six Bloom predictors while retaining approximately comparable marginal scales and the same direct coefficients. Seven conditions were repeated in this architecture: one baseline, three sample-size conditions, and three prevalence conditions. Finally, six fixed-intercept effect-magnitude conditions (multipliers 0, 0.25, 0.5, 0.75, 1.0, and 1.5) were retained to show the total system consequence when prevalence was allowed to change after modifying the Bloom effect. The complete design therefore

comprised 16 correlated prevalence-standardized conditions + 7 independent-predictor conditions + 6 fixed-intercept conditions = 29 conditions.

2.12. Implementation Sanity Check

A deliberately high-signal implementation sanity check was specified solely to test whether the model–PFI/oracle–PFI matching pipeline could recover an unambiguous leading variable when recovery should be straightforward. This condition used independent Bloom predictors, $n = 5000$, an expected event prevalence of 35%, and an eight-fold multiplier on the programmed Bloom effects. The same variable names, held-out permutation logic, oracle calculation, and ranking code used in the main analysis were retained. The purpose of this condition was limited to excluding a coding or variable-alignment failure as an explanation for weak recovery under realistic rare-event conditions; it was not used to define any general threshold for explanation validity.

2.13. Pipeline-Robustness Analysis

A separate robustness condition repeated the correlated baseline architecture ($n = 1800$, expected prevalence 3.6%) using a substantially larger random forest (2000 trees), 50 held-out permutations per variable, and no class upsampling. All other data-generating and oracle-PFI definitions were unchanged. This analysis was designed to test whether weak explanation recovery in the main Monte Carlo experiment was primarily attributable to the computationally lighter 200-tree, five-permutation, upsampled pipeline.

2.14. Conditional Permutation Importance with a Matched Conditional Oracle

Marginal permutation importance and conditional importance are different estimands. Benchmarking a conditional estimate against a marginally permuted oracle would therefore reproduce exactly the estimand mismatch described in Section 2.10. For this reason, the oracle was recomputed under the same conditional intervention, rather than reused from the marginal analysis.

For each variable X_j, two interventions were applied to the same held-out observations, within the same replicate, using the same fitted random forest:

- **Marginal.** X_j was replaced by a random permutation of its own held-out values, identical to the procedure used throughout the main Monte Carlo experiment.
- **Conditional.** X_j was replaced by a draw from $P(X_j \mid X_{-j})$. The conditional distribution was estimated on the training partition only, by regressing X_j on every other model predictor and drawing replacement values as the fitted value plus Gaussian noise with the residual standard deviation. This is the conditional-randomization sampler of Candès et al. [38], the construction underlying knockoff-based conditional predictive impact [39]. Draws were rounded and clamped to the observed support. The sampler exploits the approximately linear-Gaussian structure of the synthetic Bloom cascade. However, because the conditioning set contains all other model predictors, including correlated and derived variables, and because replacement draws are restricted to observed support, it should be regarded as a working conditional approximation rather than an exact closed-form conditional distribution. In an empirical dataset, the conditional model and its encoding assumptions would themselves require validation.

Model and oracle importances were computed from the same replacement values under both interventions, with the loss definitions of Section 2.10 unchanged. Because both interventions were applied within a single replicate, to a single fitted model and a single data split, marginal and conditional recovery were compared with paired inference: McNemar's test for top-1 recovery and for true null-first-ranking, and a paired interval for the difference in oracle–model rank agreement.

Three conditions were examined with 2000 replicates each: the correlated baseline ($n = 1800$, expected prevalence 3.6%), the correlated 10% prevalence condition, and the high-signal implementation check of Section 2.12. The analysis reuses the random-number seeds of the main experiment so that the marginal arm is designed to reproduce the corresponding main-experiment replicates. Any difference between arms is therefore attributable to the intervention rather than to the data, the split, or the fitted model. Exact replicate-level agreement was verified and is reported with the Results.

The high-signal condition was regenerated under independent seeds rather than reusing those of Section 2.12, so it constitutes an independent replication of that check.

2.15. Subgroup Analyses

Random-forest out-of-fold predictions were descriptively examined by contractor type, education level, and language match. Subgroups with fewer than 15 observed incidents were pre-labelled exploratory. These analyses were not used to infer subgroup-specific causal effects.

2.16. Software and Reproducibility

All analyses were conducted in R. The reproducibility package for this occupational-safety study includes the DGM, the primary analysis script, both competency datasets, half-effect and null-effect scenarios, codebooks, explicit Bloom ground-truth, true-null-control definitions, the parameter-provenance table, the final Monte Carlo script, replicate-level output, and the condition-level Monte Carlo summary. Fixed seeds are recorded in the scripts.

3. Results

3.1. Primary Population and Structural Fidelity

The primary dataset contained 1800 workers and 62 simulated incidents (3.44%). With approximately 25 candidate predictors entering the multivariable modelling pipeline, this corresponds descriptively to only about 2.5 observed events per candidate predictor, illustrating the limited outcome information available despite the apparently large sample. The six Bloom dimensions retained differentiated distributions and moderate correlations, consistent with a non-deterministic cascade rather than duplicate measurements (Supplementary Table S1). All pre-specified structural checks were in the programmed direction: fatigue correlated positively with latent incident probability ($r = 0.295$, $p < 0.001$); Bloom *apply* correlated negatively with incident probability ($r = -0.349$, $p < 0.001$); supervision correlated positively with procedure adherence ($r = 0.148$, $p < 0.001$); and language match correlated positively with Bloom *understand* (Spearman $\rho = 0.401$, $p < 0.001$). These checks confirm that the generated occupational system behaved as intended; they do not establish explanation fidelity.

3.2. Discrimination: Unitary Versus Bloom-Expanded Representation

Table 2 reports repeated nested cross-validation. Ridge regression achieved AUC 0.689 for v1 and 0.716 for v2. Random forest achieved AUC 0.655 for v1 and 0.644 for v2. PR-AUC values were low, as expected for a 3.44% event outcome.

The paired comparison did not provide evidence of a clear discrimination advantage for the Bloom-expanded representation. For ridge regression, $\Delta AUC = 0.027$ (95% CI −0.0068 to 0.0646; $p = 0.131$). For random forest, $\Delta AUC = -0.011$ (95% CI −0.0578 to 0.0352; $p = 0.626$).

Table 2. Primary nested cross-validation discrimination and paired v2–v1 comparison.

Dataset	Model	PR-AUC	AUC	95% CI	ΔAUC v2–v1
v1 unitary	Ridge LR	0.074	0.689	0.616–0.757	—
v2 Bloom	Ridge LR	0.075	0.716	0.647–0.776	0.027 ($p = 0.131$)
v1 unitary	Random forest	0.055	0.655	0.594–0.720	—
v2 Bloom	Random forest	0.053	0.644	0.578–0.708	−0.011 ($p = 0.626$)

AUC confidence intervals and paired differences used 2000 worker-level bootstrap resamples. The 95% CI for ridge ΔAUC was −0.0068 to 0.0646; for random forest, −0.0578 to 0.0352.

3.3. Primary-Sample Sensitivity Analyses

The prevalence-standardized falsification scenarios confirmed that the programmed direct Bloom component contributed to the simulated predictive system, but estimates from the single primary realization were imprecise. Random-forest AUC was 0.545 (95% CI 0.468–0.616) after halving the direct Bloom effect and 0.560 (95% CI 0.488–0.633) after setting the direct Bloom effect to zero.

Calibration evaluated without class upsampling yielded Brier skill scores close to zero across all model–dataset combinations (Supplementary Table S2), indicating little improvement in probabilistic accuracy over a prevalence-only reference in this sparse-event realization. Calibration slopes were also unstable; the ridge v1 slope of 3.626 is consistent with marked under-dispersion of predicted risks, while the v2 random-forest slope was 0.725. These results reinforce the limited probability information available from only 62 events.

Predictive ablation produced a full-v2 random-forest AUC of 0.644 and an AUC of 0.587 after removing both Bloom dimensions and the four behavioral descendants. The paired difference was 0.056 (95% CI −0.0035 to 0.1203; $p = 0.072$). The interval is wide and includes zero; the analysis therefore provides an imprecise estimate of the incremental predictive contribution of the observed competency-related block rather than evidence for a causal effect (Supplementary Table S3).

Domain-level held-out importance was small relative to between-fold variability, and no stable preventive hierarchy was supported (Supplementary Table S4). In the single primary realization, *remember* had the largest estimated Bloom PFI despite a zero direct coefficient, and one true-null control ranked above several Bloom dimensions (Supplementary Table S5). These single-realization rankings are reported only as illustrations of instability; the main inference is based on the Monte Carlo analysis below.

3.4. Oracle-PFI Monte Carlo Signal Recovery

The oracle-PFI benchmark is the principal explanation-fidelity result. Under the baseline prevalence-standardized correlated architecture ($n = 1800$, expected prevalence 3.6%), mean random-forest AUC was 0.689 (SD 0.064). Despite this moderate discrimination, agreement between model and oracle Bloom importance was weak: mean Spearman $\rho = 0.048$ (SD 0.445; IQR −0.261 to 0.377; 53.7% of replicates had $\rho > 0$) and mean Kendall $\tau = 0.037$.

Importantly, weak model recovery did not arise because the oracle target itself lacked a dominant top variable. *Bloom evaluate* was the oracle top-ranked Bloom variable in 1878 of 2000 baseline replicates (93.9%). It was also the Bloom dimension with the largest programmed absolute direct coefficient in the DGM. This rank-1 coincidence makes the baseline benchmark especially transparent, while not implying that coefficient magnitude and PFI are equivalent estimands. Nevertheless, the fitted random forest recovered the oracle top variable as model top-1 in only 20.7% of replicates (95% CI 18.9–22.5%) and within the top two in 37.1% (95% CI 35.0–39.2%). Operationally, the question in most baseline replicates was therefore simple: would the fitted model identify *Bloom evaluate* as the leading variable? It did so only about one time in five. With six Bloom dimensions, the corresponding nominal uniform-ranking references are 16.7% and 33.3%. The difference from the uniform reference is statistically detectable with 2000 replicates but small in practical terms.

A true-null control ranked first among the eight evaluated variables in 26.9% of baseline replicates (95% CI 25.0–28.8%), close to the 25% nominal uniform-ranking reference for two null controls among eight candidates. This does not mean that null variables are systematically favored; rather, it shows that under sparse events the fitted importance ordering contains too little stable signal to separate genuine predictors from irrelevant controls reliably.

The baseline Monte Carlo AUC (0.689) should not be interpreted as inconsistent with the v2 random-forest AUC of 0.644 in the primary dataset. The 0.644 estimate comes from repeated nested cross-validation of one realized population containing 62 events, whereas 0.689 is the mean across 2000 independently regenerated baseline populations analyzed with a computationally lighter single 70/30 split and a fixed 200-tree random forest. The two quantities therefore answer different questions: one estimates performance in a single sparse realization under nested resampling, and the other characterizes average behavior of the specified pipeline across regenerated populations.

3.5. Sample Size, Prevalence, and Effect Magnitude

Table 3 reports the selected prevalence-standardized conditions discussed below.

Table 3. Selected prevalence-standardized Monte Carlo results (2000 replicates per condition).

Architecture/Sweep	Condition	Mean AUC	Mean Oracle ρ	Top-1 Recovery (95% CI)	Null Top-1 (95% CI)
Correlated baseline	n = 1800, prev 3.6%	0.689	0.048	20.7% (18.9–22.5)	26.9% (25.0–28.8)
Correlated n	600	0.654	0.002	17.5% (15.8–19.2)	26.5% (24.6–28.5)
Correlated n	1200	0.676	0.030	18.8% (17.1–20.5)	22.9% (21.1–24.7)
Correlated n	3600	0.709	0.060	18.7% (17.1–20.5)	25.4% (23.6–27.4)
Correlated prevalence	2%	0.663	0.048	18.5% (16.8–20.2)	24.6% (22.7–26.5)
Correlated prevalence	6%	0.705	0.098	20.4% (18.7–22.2)	22.4% (20.7–24.3)
Correlated prevalence	10%	0.715	0.192	23.1% (21.3–24.9)	15.9% (14.4–17.6)
Correlated effect	1.5×	0.715	0.071	20.5% (18.7–22.3)	21.6% (19.9–23.5)
Independent baseline	n = 1800, prev 3.6%	0.655	0.063	19.1% (17.4–20.8)	24.3% (22.5–26.3)
Independent prevalence	10%	0.686	0.126	21.8% (20.0–23.6)	16.2% (14.7–17.9)

Top-1 recovery is the proportion of replicates in which the oracle most important Bloom variable was recovered as model rank 1. "Null top-1" is the proportion in which either of two true-null controls ranked first among eight candidate variables. Nominal uniform-ranking reference values are 16.7% and 25%, respectively. Confidence intervals are binomial 95% intervals across Monte Carlo replicates.

Increasing sample size from 600 to 3600 at fixed 3.6% prevalence increased mean AUC from 0.654 to 0.709. Explanation recovery improved much less consistently: mean oracle–model Spearman correlation moved from 0.002 to 0.060, while top-1 recovery was 17.5%, 18.8%, and 18.7% for $n = 600$, 1200, and 3600, respectively.

Increasing event prevalence produced a clearer signal. At 2%, 6%, and 10% prevalence, mean AUC was 0.663, 0.705, and 0.715; mean oracle–model Spearman correlation was 0.048, 0.098, and 0.192; and top-1 recovery was 18.5%, 20.4%, and 23.1%. Top-2 recovery reached 42.4% (95% CI 40.2–44.5%) at 10% prevalence. The probability that a true-null control ranked first declined to 15.9% at 10% prevalence. Figure 2 shows the prevalence gradient in top-1 recovery.

Increasing the direct-effect multiplier to 1.5 raised mean AUC to 0.715, but mean oracle–model ρ remained only 0.071 and top-1 recovery was 20.5%. Therefore, a stronger programmed signal improved overall prediction more clearly than it improved the reliability of the variable-importance ranking. Figure 3 summarizes the separation between discrimination and explanation fidelity across all 16 prevalence-standardized correlated Monte Carlo conditions.

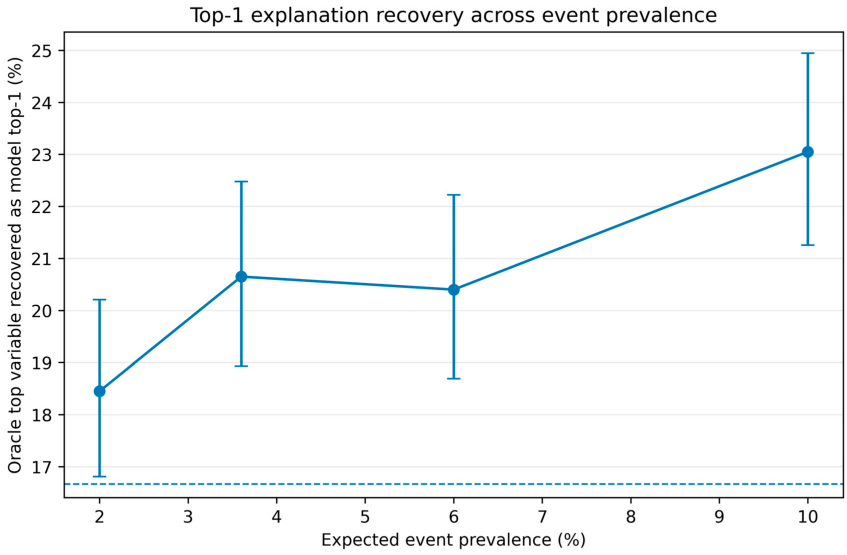


Figure 2. Recovery of the oracle top-ranked Bloom variable as model top-1 across event-prevalence conditions. Error bars show 95% binomial confidence intervals across 2000 Monte Carlo replicates. The dashed horizontal line denotes the 16.7% nominal uniform-ranking reference for one top variable among six Bloom dimensions.

3.6. Implementation Sanity Check and Pipeline Robustness

The deliberately high-signal implementation sanity check behaved as expected. Across 2000 independently generated replicates, realized prevalence averaged 35.0%, mean AUC was 0.855 (SD 0.010), and mean oracle–model Spearman agreement was $\rho = 0.964$ (SD 0.037; IQR 0.928–0.986). The oracle top variable was recovered as model top-1 in 83.4% of replicates (95% CI 81.7–85.0%) and within the top two in 98.5% (95% CI 97.8–98.9%). Neither true-null control ranked first in any replicate. Thus, although top-1 recovery was not perfect, the same PFI/oracle matching pipeline showed strong rank agreement and near-complete top-2 recovery when the signal was deliberately made easy to recover, arguing against a systematic variable-alignment or benchmark-implementation failure.

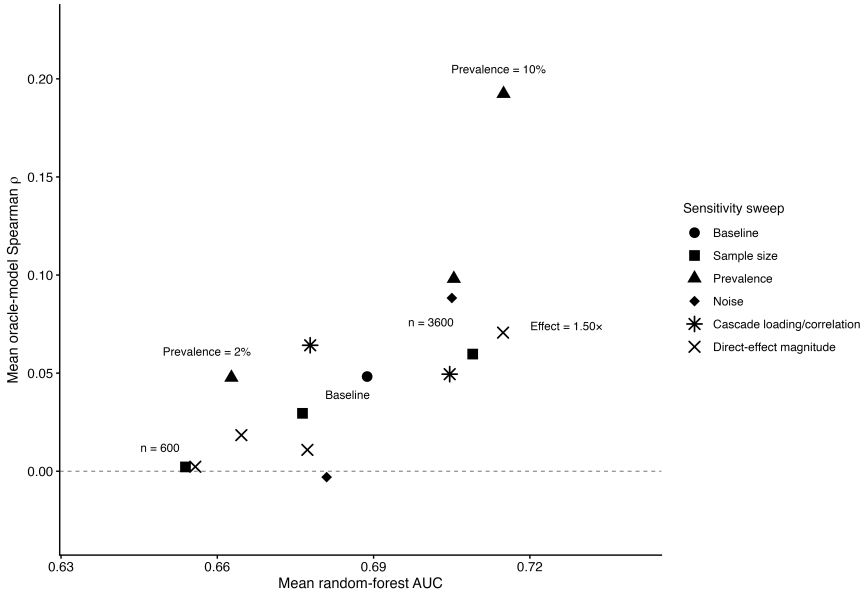


Figure 3. Predictive discrimination versus explanation fidelity across all 16 prevalence-standardized correlated Monte Carlo conditions. Marker shape identifies the sensitivity sweep; selected anchor conditions are labelled to preserve readability. Higher AUC did not imply a commensurate increase in oracle–model rank agreement. Complete numerical results for all 16 conditions are reported in Supplementary Table S8.

The computational robustness analysis reached a different conclusion under the original rare-event correlated architecture. With a 2000-tree forest, 50 held-out permutations per variable, and no class upsampling, mean AUC was 0.698 (SD 0.065) and mean oracle–model Spearman agreement was $\rho = 0.152$ (SD 0.442; IQR −0.145 to 0.522). Oracle top-1 recovery was 20.3% (95% CI 18.6–22.2%), top-2 recovery was 38.4% (95% CI 36.2–40.5%), and a true-null control ranked first in 15.8% (95% CI 14.3–17.5%). Relative to the main baseline pipeline, the more intensive specification modestly improved average rank agreement and reduced null-first frequency, but it did not materially improve recovery of the oracle top variable (20.3% versus 20.7%). The central explanation-fidelity finding therefore does not depend on the lighter forest, limited permutation count, or class upsampling used in the main Monte Carlo experiment.

3.7. Independent-Predictor Comparator Architecture

Removing the Bloom cascade and shared latent dependence did not restore strong recovery. Under the independent baseline architecture, mean AUC was 0.655, mean oracle–model Spearman $\rho = 0.063$, top-1 recovery was 19.1%, and top-2 recovery was 37.2%. At 10% prevalence, recovery improved modestly ($\rho = 0.126$, top-1 21.8%, top-2 39.3%), while true-null top-1 frequency fell to 16.2%.

The weak recovery observed in the correlated architecture therefore cannot be attributed solely to the Bloom cascade. Under the tested random-forest/PFI pipeline, sparse events and limited predictive signal appear to be important constraints even when the candidate predictors are independent.

3.8. Conditional Importance Does Not Restore Ranking Fidelity

Under the correlated baseline, conditional importance did not improve recovery of the oracle top variable. Top-1 recovery was 20.4% (95% CI 18.6–22.2%) under the conditional intervention and 20.7% (95% CI 18.9–22.5%) under the marginal intervention; the paired difference was −0.3 percentage points (177 replicates recovered only under the conditional

intervention, 183 only under the marginal; McNemar $p = 0.79$). Mean oracle–model rank agreement was effectively identical, with a paired difference in Spearman ρ of +0.001 (95% CI −0.018 to +0.019). The interval is narrow, so this is a precise null rather than an underpowered comparison.

The conditional intervention did, however, change one quantity substantially. The probability that a true-null control ranked first among the eight evaluated variables fell from 26.9% to 22.9%. The paired comparison was clearly directional: 232 replicates showed a null-first selection under the marginal intervention only, against 153 under the conditional intervention only (McNemar $p < 0.001$). This is the behavior predicted by work showing that unrestricted permutation forces extrapolation into regions of low data density and redistributes apparent importance [12]. Conditional sampling mitigated that specific pathology without improving identification of the genuinely most important variable.

At 10% prevalence the pattern was similar for top-1 recovery (24.0% versus 23.1%; McNemar $p = 0.38$) and for null-first selection (14.9% versus 15.9%; $p = 0.27$), but mean rank agreement was slightly *lower* under the conditional intervention ($\rho = 0.147$ versus 0.192; paired difference −0.046, 95% CI −0.064 to −0.027). Conditional importance was therefore not a uniform improvement on the marginal estimator in this setting. Two further results bear on how these comparisons should be read. First, the relative separation of the oracle target—the gap between the first- and second-ranked oracle importances, expressed as a fraction of the first—was larger under the conditional intervention in both correlated conditions (baseline median 0.25, IQR 0.21–0.30, versus 0.15, IQR 0.09–0.21; 10% prevalence median 0.25 versus 0.16). The two gaps describe different estimands and so do not share a strict common scale, but the comparison indicates that weaker conditional recovery cannot be attributed to a less well-separated target.

Second, in the high-signal implementation, the Bloom predictors are independent by construction, so the conditional and marginal interventions coincide in distribution. They did: mean ρ was 0.966 and 0.965 respectively, with a paired difference of +0.0004 (95% CI −0.001 to +0.002), top-1 recovery was 85.2% and 84.6% (McNemar $p = 0.47$), neither true-null control ranked first in any replicate under either intervention, and the oracle separation distributions were indistinguishable (median 0.125 versus 0.126). This provides an internal validity check on the conditional sampler: where there is no predictor dependence to exploit, it reduces to the marginal procedure.

That same condition also quantifies why first-rank recovery is intrinsically demanding even when overall agreement is near-perfect. With mean $\rho = 0.966$, the oracle top variable exceeded the second-ranked variable by a median of only 12.6% of its own magnitude (IQR 9.0–15.8%). A target separated this narrowly at the first rank will not be recovered in every replicate even by a model that orders the remaining variables almost perfectly, which accounts for top-1 recovery of 84.6% alongside near-complete top-2 recovery.

This independently seeded replication of the high-signal check produced 84.6% top-1 recovery (95% CI 82.9–86.2%), consistent with the 83.4% obtained in the original run reported in Supplementary Table S6. Full results appear in Supplementary Table S7.

3.9. High-Exposure Descriptive Contrast and Subgroups

Among workers with high chemical exposure, the lowest quartile of Bloom *apply* contained 15 incidents among 151 workers (9.9%), compared with 8 incidents among 156 workers (5.1%) in the highest quartile. Mean latent incident probability was 0.088 versus 0.024. Because the DGM itself encodes a protective *apply* coefficient, this contrast is a descriptive property of the simulation and should not be interpreted as empirical evidence of an interaction in real workers.

Subgroup discrimination varied widely. Several strata contained fewer than 15 events, and even larger strata produced broad confidence intervals. These analyses are therefore retained as exploratory robustness information rather than a basis for subgroup-specific recommendations.

4. Discussion

4.1. Main Finding for OSH Practice: A Useful Risk Score Is Not Necessarily a Trustworthy Priority List

The central occupational-safety finding is that the two outputs a practitioner may receive from the same AI system can have very different evidential quality. Under the baseline Monte Carlo condition, the random forest achieved mean AUC 0.689, indicating non-trivial ability to order higher-risk and lower-risk observations. Yet the corresponding importance ranking recovered the oracle ordering only weakly: mean Spearman agreement was 0.048, and the oracle top variable was identified as model top-1 in only 20.7% of simulations.

Two additional analyses narrow the interpretation of this result. First, under a deliberately easy high-signal condition, the same implementation produced mean oracle–model $\rho = 0.964$, 83.4% top-1 recovery, 98.5% top-2 recovery, and no null-first selections, showing that the benchmarking chain can recover a strong signal when one is present. Second, under the rare-event baseline architecture, increasing the forest to 2000 trees, using 50 permutations per variable, and removing upsampling left top-1 recovery essentially unchanged at 20.3%. These checks make it less plausible that the principal finding is an artefact of variable-name alignment, an incapable benchmark implementation, or the computationally lighter random-forest specification used for the 58,000-fit Monte Carlo design.

For practice, this distinction is more important than whether one algorithm marginally outperforms another. A screening model may still help identify workers, tasks, shifts, or circumstances that deserve closer review. That does not mean that the same model's feature-importance table is sufficiently reliable to decide whether the organization should invest first in training, supervision, procedures, staffing, or another preventive measure.

Accordingly, the study does not conclude that random forests or PFI are unsuitable for occupational safety. It concludes that predictive usefulness and ranking fidelity require separate evaluation when the ranking is intended to influence a preventive decision. In this setting, model accuracy alone is insufficient evidence for treating a ranked factor as the organization's next preventive priority.

4.2. What the Fidelity Metrics Mean for a Safety Professional

The quantities used above are unfamiliar outside methodological work, so it is worth stating what each one would mean to someone reading a dashboard.

Top-1 recovery is the probability that the factor the system places at the top of its list is in fact the factor that matters most for prediction. Under the baseline condition, this was approximately one in five. Put differently, across repeated comparable datasets, roughly four in five fitted rankings failed to place the oracle top variable first. The fitted output alone did not identify which rankings had recovered the correct first-ranked variable.

Rank agreement describes the whole ordering rather than its first position. A Spearman correlation of 0.048 is close to the value expected if the ordering had been produced at random. It does not mean the ranking is actively misleading; it means the ordering carries very little information about the true ordering.

True null-first-ranking is the most concrete of the three. Two variables in the system were constructed with no pathway of any kind to the outcome: they cannot matter, by design. In the baseline condition, one of them was nevertheless ranked above all six genuine

competency dimensions in roughly one replicate in four. A dashboard producing that output would be presenting a variable with no causal or predictive role as the organization's leading risk factor, and nothing in the output would signal this.

Separation of the target explains why first-rank recovery is demanding even when a model performs well. When the two most important factors are genuinely close, no estimator can reliably distinguish them, and an ordering that puts either one first is equally defensible. This is a property of the system being studied, not a defect of the analysis—but it does mean that a list ordering factors 1 to 6 conveys more precision than the underlying data support.

Taken together, these are not arguments against using predictive models in prevention. They are arguments for reading a ranked list as a set of candidates for investigation rather than as an ordering of priorities.

4.3. Why an Occupational-Safety Ranking Needs an Independent Fidelity Check

A practical difficulty in evaluating an AI ranking is that a plausible ordering may look convincing even when there is no known reference against which to judge it. The synthetic design used here makes such a check possible because the DGM is fully known. We therefore calculated oracle-PFI from the DGM itself and compared the fitted-model–PFI with the same predictive quantity under the same held-out permutations.

This like-for-like comparison is important in the present occupational application because raw programmed coefficients are not equivalent to PFI. A coefficient describes one component of the DGM, whereas PFI depends on the predictor distribution, dependence structure, loss function, and the fitted prediction rule. The oracle benchmark therefore serves here as a validation device for the ranking shown to a hypothetical practitioner, rather than as a claim that one universal explanation benchmark should be used in every AI application.

The persistence of weak recovery under this occupationally grounded benchmark means that the concern cannot be dismissed as a mismatch between coefficient magnitude and model importance. In practical terms, the fitted ranking often failed to recover a predictive hierarchy that was known to exist by construction.

4.4. Correlation Is Not the Only Explanation

Correlated predictors are a recognized challenge for permutation methods [11]. The Bloom case deliberately incorporates correlation because cognitive dimensions are unlikely to be statistically independent in real training data. However, the independent comparator architecture showed similarly weak recovery at baseline. Removing the cascade did not transform the PFI ranking into a reliable recovery mechanism.

This does not imply that correlation is irrelevant. Different dependence structures can still affect individual importance estimates. Section 4.5 reports a direct test of one dependence-aware estimator; other conditional and dependence-aware importance methods deserve further study. It does mean that, in the present experiment, correlation alone is insufficient to explain the observed instability. Event scarcity and weak effective signal appear to play a substantial role.

4.5. Conditional Importance Addresses Extrapolation, Not Event Scarcity

A natural response to weak ranking recovery is that the wrong importance estimator was used. Marginal permutation is known to be vulnerable when predictors are dependent, because permuting a variable independently of the others generates feature combinations that are rare or implausible in the original data [12,13]. We therefore repeated the benchmark with conditional permutation importance, recomputing the oracle under the same conditional intervention so that the estimate and target remained the same quantity.

The result separates two things that are easily conflated. Conditional sampling did reduce one specific failure: the frequency with which a variable with no pathway whatsoever to the outcome was ranked first fell from 26.9% to 22.9% in the baseline condition, with a clearly directional paired comparison. That is the extrapolation artefact behaving as the methodological literature predicts, and being partly corrected.

However, the correction did not transfer to the quantity that matters for preventive prioritization. Recovery of the genuinely most important variable was unchanged (20.4% versus 20.7%), with a paired confidence interval narrow enough to exclude any practically meaningful improvement, and overall rank agreement was slightly worse at 10% prevalence. A better-behaved estimator did not produce a more trustworthy priority list.

For OSH practice, the implication is specific and worth stating plainly. If an analyst is told that a ranking is unreliable, the natural reaction is to reach for a more sophisticated explanation method. In the conditions examined here, replacing marginal PFI with the dependence-aware conditional estimator did not restore top-rank recovery. Within these random-forest pipelines, limited outcome information remained an important constraint even after the importance intervention was changed. This finding should not be interpreted as evidence that estimator choice is unimportant in other model classes or explanation frameworks.

This also sharpens what the present study does and does not show. The weak recovery reported throughout is not an artefact of marginal permutation, since an estimator designed specifically to handle predictor dependence, benchmarked against a matched oracle, behaved the same way. At the same time, the finding remains bounded by the two estimators actually tested and by the synthetic rare-event conditions specified here.

4.6. Rare Incidents Limit What Practitioners Can Infer from Large Safety Datasets

One of the most practically relevant findings is the different behavior of sample size and event prevalence. Increasing n from 600 to 3600 improved AUC substantially, but top-variable recovery remained near 18–19%. Increasing prevalence from 2% to 10%, by contrast, produced a clearer increase in oracle–model rank agreement and reduced the frequency with which true-null controls appeared at the top.

For occupational-safety research, this distinction is critical. Safety datasets are often "large" in number of workers, shifts, inspections, or exposure records but contain relatively few actual injury or incident events. Prediction-model methodology emphasizes that outcome-event information, not merely the total number of records, constrains model development and precision [34]. The present sweeps additionally show that absolute event count and event prevalence are not interchangeable: increasing sample size at fixed 3.6% prevalence increased the expected number of events but produced little change in top-1 recovery, whereas increasing prevalence at fixed n produced clearer gains in rank agreement. In this logistic synthetic system, explanation recovery is therefore best understood as depending jointly on event count, prevalence-dependent signal geometry, predictor dependence, and effect strength rather than on a single event-count threshold.

The study does not establish a universal minimum number of incidents. No such threshold can be inferred from these simulations. The practical conclusion is instead conditional: explanation methods should be validated under event counts, prevalence, predictor dependence, signal strength, and model family similar to those in the intended application.

4.7. Bloom Disaggregation: Useful for Training Design, Not Proven as a Superior Predictor

The Bloom-expanded representation did not clearly outperform the unitary SDS-competency score. This is noteworthy because the outcome DGM explicitly contained Bloom

direct terms and was therefore favorable to the richer representation. Ridge regression showed a small positive ΔAUC, but its confidence interval included zero; random forest showed a small negative difference, also with an interval spanning zero.

This finding should not be interpreted as evidence against Bloom's Taxonomy as a pedagogical framework. A multidimensional assessment may still be valuable for diagnosing learning needs, designing scenarios, or structuring competency development. The present study addresses a narrower question: whether six correlated Bloom indicators add demonstrable predictive discrimination for a sparse simulated incident outcome beyond a unitary competency proxy. Under the tested conditions, no clear advantage was demonstrated.

4.8. Why Domain Rankings Should Not Be Converted Directly into Preventive Priorities

Domain-level held-out importance did not support a stable ordering in which organizational conditions were clearly the leading predictive domain. Mean held-out importance was highest for intermediate behaviors and Bloom competency, but between-fold variability was large, and the ordering was unstable.

This is more than a statistical footnote. If an OSH professional were shown the same output in a commercial dashboard, ranking the domains from 1 to 6 would create an impression of precision that the analysis does not support. A prevention plan should not reallocate resources from supervision to training, or vice versa, merely because one domain appears slightly higher in a single importance table.

The appropriate use of such a ranking is as a prompt for investigation: inspect the underlying exposure, organizational, behavioral, and competency data; determine whether the factor is modifiable; assess whether the association is plausible and independently supported; and test whether the ranking is stable under resampling, alternative models, or held-out periods.

4.9. A Practitioner-Facing Validation Rule for AI-Ranked Risk Factors

The study supports a deliberately conservative rule for occupational-safety practice: *use a risk score to decide where to look only if its predictive performance has been validated, and use a factor ranking to decide what to prioritize only if the ranking itself has also been validated*. These are separate checks.

Before an AI-derived ranking is used to justify a training redesign, additional supervision, targeted inspection, staffing change, or reallocation of preventive resources, the OSH professional should examine at least four issues. First, how much outcome information supports the ranking, considering both event count and prevalence? Second, were predictive performance and the explanation evaluated on observations not used for model fitting or tuning? Third, was the ranking challenged against an independent benchmark, negative control, synthetic reference, or temporal/site replication? Fourth, is the ranking stable under resampling, repeated fitting, or reasonable alternative model specifications?

Each item in Table 4 is motivated either by a result reported above or by the corresponding validation requirement needed to interpret that result safely. Item 1 follows from the contrasting sample-size and prevalence sweeps: outcome information depends on both event count and prevalence rather than on dataset size alone. Item 2 is the foundational out-of-sample requirement used throughout the study: both predictive performance and importance were evaluated on observations not used to fit the explained model. Item 3 follows from the oracle benchmark itself: a plausible-looking ordering was recovered correctly about one time in five, and nothing in the output distinguished the successful replicates from the unsuccessful ones. Item 4 follows from the marked replicate-to-replicate instability of the ranking under sparse outcomes. Item 5 follows from the comparison between the correlated and independent architectures, and from the conditional-importance analysis,

which reduced spurious first-ranking of true-null controls without improving recovery of the correct variable. Item 6 follows from the ablation analysis, where removing the competency block changed out-of-sample performance only imprecisely. Item 7 follows from the design of the mechanism, in which two dimensions with zero direct coefficients could still rank highly as proxies for correlated information. Item 8 reflects the overall position of the study rather than any single result.

Table 4. Preliminary appraisal checklist, derived from the simulation results reported here, for use before an AI-derived risk-factor ranking is allowed to influence preventive prioritization. The checklist has not been independently validated.

#	Question for the Practitioner	Why It Matters
1	How much outcome information supports the model, considering both the number of events and event prevalence?	Large datasets can still provide weak ranking information when outcomes are rare, and risk gradients are compressed.
2	Were predictive performance and the explanation evaluated on workers, periods, or sites not used for model fitting or hyperparameter selection?	In-sample performance and explanation can be optimistically biased.
3	Was the importance method challenged against a known benchmark, negative control, synthetic reference, or held-out temporal/site replication?	A plausible-looking ranking does not demonstrate that the method recovers the relevant predictive structure.
4	Is the ranking stable under resampling, repeated fitting, reasonable alternative models, and small changes in the data?	A preventive priority should not depend on a fragile rank ordering.
5	Are highly correlated variables or domains present, and if so, was dependence addressed explicitly?	Marginal permutation can redistribute or mask importance among correlated predictors.
6	Does removing or withholding the highlighted factor meaningfully reduce out-of-sample predictive performance?	If performance barely changes, the factor may be replaceable by correlated information and should not be treated as uniquely decisive.
7	Is the output explicitly predictive, or is a causal/intervention claim being made?	Predictive importance does not establish that intervening on the factor will reduce incidents.
8	Is the ranking interpreted alongside conventional risk assessment, exposure evidence, legal duties, worker consultation, and professional judgement?	XAI should support, not replace, the preventive decision process.

Table 4 is offered as a preliminary proposal derived from these simulation results, not as a validated instrument. Its items have not been tested for inter-rater agreement, for discrimination between trustworthy and untrustworthy rankings in real settings, or for any effect on preventive decisions. It is a conceptual tool intended to structure professional judgement; it is not a regulatory instrument, and it does not replace conventional risk assessment. Its purpose is to prevent a common interpretive shortcut: moving directly from "the model predicts reasonably well" to "the highest-ranked variable should receive the highest preventive priority". Empirical validation of the checklist is itself a task for future research.

A practical governance rule also follows. If an AI ranking materially influences an OSH decision, the organization should retain the model version, evaluation dataset, explanation method, number of observed events, stability or uncertainty analysis, and the professional rationale for the action taken. This documentation creates an auditable preventive decision process rather than a record that merely states that "the AI identified factor X as the main risk."

4.10. Calibration and the Danger of Pipeline Artefacts

The calibration analysis illustrates a separate but related lesson. Class upsampling changes the outcome prevalence seen during fitting and can distort absolute probability estimates if those probabilities are interpreted without adjustment. For that reason, calibration in the present study was evaluated in a separate nested cross-validation pipeline without class upsampling. Under that target-prevalence-preserving pipeline, Brier skill scores were close to zero rather than showing the extreme degradation that can arise when class-balanced fitting probabilities are interpreted directly.

For applied OSH analytics, this example reinforces the need to distinguish model behavior from preprocessing effects. Class balancing may help discrimination but alter probability interpretation. If an AI system presents absolute incident probabilities, calibration must be checked on data retaining the target population prevalence.

4.11. Structural Fidelity Versus Explanation Fidelity

All four structural checks confirmed that the simulation behaved as programmed. Yet explanation recovery remained weak. This demonstrates why structural fidelity and explanation fidelity should be reported separately in synthetic-data research.

Structural fidelity asks whether the data generator contains and expresses the intended relationships. Explanation fidelity asks whether a downstream analytical pipeline discovers the relevant predictive structure. A synthetic dataset can satisfy the first condition without the fitted explanation satisfying the second.

This distinction is useful beyond the present Bloom case. Any synthetic benchmark intended to validate an XAI method should define the explanation target explicitly and, where possible, calculate an oracle version of that same target.

4.12. Limitations

Several limitations constrain interpretation. First, the study uses synthetic rather than empirical occupational data. The DGM is theory-informed but not a calibrated model of a specific oil operation, workforce, or national incident system. Coefficients should therefore be interpreted as explicit simulation parameters, not empirical effect estimates.

Second, several DGM parameters, particularly Bloom cascade loadings and some workforce-composition probabilities, remain modelling assumptions. Published sources were used as directional or order-of-magnitude anchors where possible, but the simulation was not fitted to an external individual-level dataset.

Third, the main Monte Carlo analysis uses an untuned 200-tree random forest on one 70/30 split per replicate rather than repeated nested cross-validation within every replicate. This was necessary to keep 58,000 independent simulation fits computationally reproducible. A dedicated robustness analysis using 2000 trees, 50 permutations per variable, and no upsampling produced essentially the same top-1 recovery (20.3% versus 20.7%), although average rank agreement improved modestly and null-first selections decreased. The Monte Carlo should therefore be interpreted as characterizing recovery patterns under the specified modelling family rather than estimating one optimally tuned model.

Fourth, the sensitivity analysis is mainly one factor at a time. Interactions such as simultaneously increasing sample size and prevalence were not exhaustively explored. A future factorial or space-filling design could identify conditions under which explanation recovery becomes acceptably reliable.

Fifth, the study evaluates two importance estimators: conventional held-out marginal permutation importance and conditional permutation importance with a matched conditional oracle, both applied to random forests. The two behaved equivalently with respect to recovery of the oracle top variable, which makes it unlikely that the central finding is an

artefact of marginal permutation specifically. However, the oracle-recovery experiment did not compare model families. Ridge logistic regression achieved higher discrimination than random forest in the primary dataset, and the present results therefore do not establish whether a lower-variance parametric model would recover the same importance hierarchy more faithfully under comparable event scarcity. Other approaches may also behave differently: SHAP variants [40], knockoff-based inference, grouped attribution [37], causal models, and estimators that target conditional rather than marginal predictive relevance are not covered by these results. The oracle benchmark developed here is intended to facilitate exactly those comparisons, since it can be recomputed under whichever intervention the estimator in question implies. The conditional sampler used here is an approximately linear-Gaussian working model tailored to the synthetic system; with empirical data, the sampler would itself be an estimation problem and a possible source of error. Accordingly, the demonstrated limitation concerns the random-forest/permutation-importance family under the specified synthetic rare-event conditions, and should not be read as a general claim about AI-based explanation methods.

Sixth, the deliberately high-signal implementation sanity check was designed only to test whether the model–PFI/oracle–PFI chain could recover an unambiguous signal under favorable conditions. Its strong rank agreement and near-complete top-2 recovery argue against a systematic alignment failure, but its 83.4% top-1 recovery also shows that even an intentionally easy random-forest/PFI setting need not yield perfect first-rank recovery. It should therefore not be treated as a universal performance threshold for PFI or other explanation methods. Likewise, the independent comparator architecture was intended only to test whether predictor dependence alone explained the main result.

Finally, subgroup analyses contain sparse strata and are exploratory. They should not be interpreted as evidence of differential model validity among education, language, or employment groups.

4.13. Future Research

The next methodological step is to extend the conditional-PFI benchmark across the full simulation grid and compare additional explanation strategies against matched targets, including alternative dependence-aware samplers, SHAP-based methods [40], grouped feature attribution [37], and simpler regression-based approaches. Such comparisons should vary event prevalence, effect strength, predictor dependence, and model family explicitly.

The substantive next step is empirical. A real occupational cohort should assess SDS competency with validated scenario-based items mapped to Bloom dimensions, while independently measuring organizational factors, exposure, and safety behaviors. The central empirical question would then be whether multidimensional competency contributes information beyond a unitary score and whether any observed importance ranking replicates across sites or time periods.

For practice, prospective validation should focus not only on whether an AI system predicts incidents but also on whether decisions made from its explanations improve preventive outcomes. A factor ranking has practical value only if it can support a reproducible and defensible decision process.

5. Conclusions

This study addresses a practical question for occupational-safety professionals: how much trust should be placed in an AI-generated ranking of "important risk factors" when that ranking may influence preventive priorities?

In the synthetic SDS/Bloom use case, a richer six-dimensional competency representation did not show a clear discrimination advantage over a unitary competency score,

despite the DGM explicitly containing Bloom-specific direct terms. Calibration results also showed that preprocessing choices can materially affect the apparent quality of absolute risk probabilities.

More importantly, the oracle-PFI Monte Carlo analysis showed that moderate predictive discrimination can coexist with weak recovery of the underlying predictive-importance hierarchy. Under the baseline rare-event condition, the random forest achieved mean AUC 0.689, while oracle–model rank agreement remained close to zero and the oracle top variable was recovered only modestly more often than a uniform-ranking reference. Increasing sample size improved discrimination more consistently than top-rank recovery, whereas higher event prevalence produced clearer gains in explanation fidelity and better separation of true-null controls. A deliberately high-signal implementation check produced strong oracle–model agreement and 98.5% top-2 recovery, whereas a substantially more intensive rare-event pipeline left top-1 recovery essentially unchanged at 20.3%. Taken together, these results support the interpretation that weak explanation recovery is a property of the rare-event correlated setting examined here rather than a simple coding or forest-specification artefact.

For OSH practice, the implication is deliberately narrow. An AI system may be useful for identifying where closer investigation is warranted while remaining unreliable about which factor should be changed first. Feature-importance rankings should therefore be treated as testable analytic hypotheses, not as automatic preventive priority lists. Before such rankings influence training, supervision, inspection, staffing, or resource allocation, practitioners should assess out-of-sample validity, event information, ranking stability, predictor dependence, negative controls or other fidelity checks, and consistency with exposure evidence, legal duties, worker consultation, and professional judgement.

The contribution of this study is consequently occupational and practice-oriented: it provides a controlled demonstration of why prediction and explanation should be evaluated separately in rare-event safety analytics and offers a preliminary appraisal checklist, derived from those results and still requiring independent validation, for professionals who must decide whether an AI-ranked factor is sufficiently trustworthy to influence prevention.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/safety12060148/s1, Table S1: Descriptive statistics for Bloom competency dimensions; Table S2: Nested cross-validation calibration without class upsampling; Table S3: Random-forest predictive ablation; Table S4: Cross-fitted held-out group permutation importance; Table S5: Individual held-out PFI and programmed direct coefficients in the primary v2 dataset; Table S6: Safety-specific implementation and pipeline-robustness checks; Table S7: Paired comparison of marginal and conditional permutation importance against matched oracles; Table S8: Complete results for all 16 prevalence-standardized correlated Monte Carlo conditions. The complete reproducibility package, including the synthetic data-generating code, analysis scripts, synthetic datasets, ground-truth definitions, Monte Carlo outputs, and replicate-level verification outputs, is openly archived in Zenodo (last date of access: 5 November 2026) at https://doi.org/10.5281/zenodo.22746852.

Author Contributions: Conceptualization, R.A.-E. and J.J.A.-G.; methodology, R.A.-E., A.D.-G. and A.G.-R.; software, R.A.-E. and A.D.-G.; formal analysis, R.A.-E., A.D.-G. and R.T.-R.; validation, R.A.-E., A.D.-G., R.T.-R. and A.G.-R.; investigation, R.A.-E.; data curation, R.A.-E., A.D.-G. and R.T.-R.; visualization, R.A.-E., A.D.-G. and R.T.-R.; writing—original draft preparation, R.A.-E.; writing—review and editing, R.A.-E., A.D.-G., R.T.-R., A.G.-R. and J.J.A.-G.; supervision, R.A.-E., A.G.-R. and J.J.A.-G.; project administration, R.A.-E. and J.J.A.-G. R.A.-E. led the conceptual, methodological, analytical, computational and writing components of the study. A.G.-R. served as corresponding author. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study used exclusively synthetic, computer-generated data and involved no human participants, identifiable human data, or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data supporting the reported analyses are synthetic. The data-generating code, analysis code, ground-truth definitions, parameter-provenance information, primary synthetic datasets, falsification scenarios, Monte Carlo outputs, implementation sanity-check outputs, and pipeline-robustness outputs are openly available in Zenodo at https://doi.org/10.5281/zenodo.22746852. Fixed seeds included in the scripts permit reproduction of the primary generated datasets and reported analyses.

Acknowledgments: During the preparation of the final manuscript, the authors used ChatGPT (GPT-5, OpenAI) solely to assist with translation of text. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CV	Cross-validation
DGM	Data-Generating Mechanism
GHS	Globally Harmonized System of Classification and Labelling of Chemicals
LR	Logistic Regression
IQR	Interquartile Range
ML	Machine Learning
OSH	Occupational Safety and Health
PFI	Permutation Feature Importance
PPE	Personal Protective Equipment
PR-AUC	Area Under the Precision–Recall Curve
RF	Random Forest
SDS	Safety Data Sheet
SD	Standard Deviation
XAI	Explainable Artificial Intelligence

References

1. 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. [CrossRef] [PubMed]
2. Obasi, I.C.; Cheng, P.; Varianou-Mikellidou, C.; Dimopoulos, C.; Boustras, G. Machine learning for occupational accident analysis: Applications, challenges, and future directions. *J. Saf. Sci. Resil.* **2026**, *7*, 100250. [CrossRef]
3. Yapi, A.; Latouche, P.; Guillin, A.; Bailly, Y. A new machine learning framework for occupational accidents forecasting with safety inspections integration. *J. Saf. Res.* **2026**, *97*, 244–260. [CrossRef] [PubMed]
4. Zailani, B.M.; Ibrahim, Y.; Ibrahim, Y.M.; Yamusa, M.A. Modeling safety compliance intentions of construction workers: An explainable machine learning approach. *Saf. Sci.* **2026**, *203*, 107335. [CrossRef]
5. Qin, R.; Cui, P.; Muhsin, J. Research Progress of Automation Ergonomic Risk Assessment in Building Construction: Visual Analysis and Review. *Buildings* **2024**, *14*, 3789. [CrossRef]
6. Gursel, E.; Madadi, M.; Coble, J.B.; Agarwal, V.; Yadav, V.; Boring, R.L.; Khojandi, A. The role of AI in detecting and mitigating human errors in safety-critical industries: A review. *Reliab. Eng. Syst. Saf.* **2025**, *256*, 110682. [CrossRef]
7. Jetha, A.; Bakhtari, H.; Irvin, E.; Biswas, A.; Smith, M.J.; Mustard, C.; Arrandale, V.H.; Dennerlein, J.T.; Smith, P.M. Do occupational health and safety tools that utilize artificial intelligence have a measurable impact on worker injury or illness? Findings from a systematic review. *Syst. Rev.* **2025**, *14*, 146. [CrossRef] [PubMed]
8. Orviz-Martínez, N.; Pérez-Santín, E.; López-Sánchez, J.I. New Trends in the Use of Artificial Intelligence and Natural Language Processing for Occupational Risks Prevention. *Safety* **2026**, *12*, 7. [CrossRef]

9. Fisher, A.; Rudin, C.; Dominici, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **2019**, *20*, 1–81. Available online: https://jmlr.org/papers/v20/18-760.html (accessed on 3 November 2026).
10. Molnar, C.; Casalicchio, G.; Bischl, B. iml: An R package for Interpretable Machine Learning. *J. Open Source Softw.* **2018**, *3*, 786. [CrossRef]
11. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed.; Christoph Molnar: Munich, Germany, 2022. Available online: https://christophm.github.io/interpretable-ml-book/ (accessed on 17 October 2026).
12. Hooker, G.; Mentch, L.; Zhou, S. Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Stat. Comput.* **2021**, *31*, 82. [CrossRef]
13. Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional variable importance for random forests. *BMC Bioinform.* **2008**, *9*, 307. [CrossRef] [PubMed]
14. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef] [PubMed]
15. Ewald, F.K.; Bothmann, L.; Wright, M.N.; Bischl, B.; Casalicchio, G.; König, G. A Guide to Feature Importance Methods for Scientific Inference. In *Explainable Artificial Intelligence, xAI 2024*; Springer: Cham, Switzerland, 2024; pp. 440–464. [CrossRef]
16. Buggineni, V.; Chen, C.; Camelio, J. Enhancing manufacturing operations with synthetic data: A systematic framework for data generation, accuracy, and utility. *Front. Manuf. Technol.* **2024**, *4*, 1320166. [CrossRef]
17. Giuffrè, M.; Shung, D.L. Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *npj Digit. Med.* **2023**, *6*, 186. [CrossRef] [PubMed]
18. Pezoulas, V.C.; Zaridis, D.I.; Mylona, E.; Androutsos, C.; Apostolidis, K.; Tachos, N.S.; Fotiadis, D.I. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 2892–2910. [CrossRef] [PubMed]
19. Rajotte, J.-F.; Bergen, R.; Buckeridge, D.L.; El Emam, K.; Ng, R.; Strome, E. Synthetic data as an enabler for machine learning applications in medicine. *iScience* **2022**, *25*, 105331. [CrossRef] [PubMed]
20. Occupational Safety and Health Administration. 29 CFR § 1910.1200: Hazard Communication Standard. Available online: https://www.osha.gov/laws-regs/regulations/standardnumber/1910/1910.1200 (accessed on 17 October 2026).
21. United Nations. *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*, 6th ed.; United Nations: New York, NY, USA, 2015.
22. United Nations. *Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*, 9th ed.; United Nations: New York, NY, USA, 2021.
23. Ho, K.; Tenkate, T. Safety Data Sheets as a Hazard Communication Tool: An Assessment of Suitability and Readability. *Saf. Health Work* **2024**, *15*, 192–199. [CrossRef] [PubMed]
24. Nicol, A.; Hurrell, A.C.; Wahyuni, D.; McDowall, W.; Chu, W. Accuracy, comprehensibility, and use of material safety data sheets: A review. *Am. J. Ind. Med.* **2008**, *51*, 861–876. [CrossRef] [PubMed]
25. Zerrouki, H.; Ghozlane, M.D.E.; Estrada Lugo, H.D.; Patelli, E. Workplace accident analysis in the Algerian oil and gas industry. *Process Saf. Prog.* **2023**, *42*, 328–337. [CrossRef]
26. De Barros Emery Trindade, R.; De Queiroz, D.V., Jr. HSE Training and Competence Development at Petrobras University. In Proceedings of the SPE International Conference on Health, Safety and Environment in Oil and Gas Exploration and Production, Rio de Janeiro, Brazil, 12–14 April 2010; SPE-126211-MS. [CrossRef]
27. Fathoni, A.Z. The role of oil and gas manpower development centre in health, safety, and environment training activities. In Proceedings of the SPE Health, Safety and Environment in Oil and Gas Exploration and Production Conference, Jakarta, Indonesia, 25–27 January 1994; SPE-27246-MS. [CrossRef]
28. Lau, H.C. Competency Development for the Next Generation of Petroleum Engineers. In Proceedings of the SPE Abu Dhabi International Petroleum Exhibition & Conference, Abu Dhabi, United Arab Emirates, 13–16 November 2017; SPE-188466-MS. [CrossRef]
29. Bloom, B.S.; Engelhart, M.D.; Furst, E.J.; Hill, W.H.; Krathwohl, D.R. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*; David McKay: New York, NY, USA, 1956.
30. Anderson, L.W.; Krathwohl, D.R. (Eds.) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*; Longman: New York, NY, USA, 2001.
31. Mashamba-Thompson, T.P.; Sartorius, B.; Stevens, F.C.J.; Drain, P.K. Experiential Bloom's Taxonomy learning framework for point-of-care diagnostics training of primary healthcare workers. *Afr. J. Lab. Med.* **2016**, *5*, 449. [CrossRef] [PubMed]
32. Tien, H.L.; Van, N.P.; Kato, T. Development of Evaluation Criteria for Training Fire Students to Enable New Rescue Roles in Vietnam. *J. Disaster Res.* **2024**, *19*, 411–419. [CrossRef]
33. Bęś, P.; Strzałkowski, P. Analysis of the Effectiveness of Safety Training Methods. *Sustainability* **2024**, *16*, 2732. [CrossRef]

34. Riley, R.D.; Ensor, J.; Snell, K.I.E.; Harrell, F.E.; Martin, G.P.; Reitsma, J.B.; Moons, K.G.M.; Collins, G.; van Smeden, M. Calculating the sample size required for developing a clinical prediction model. *BMJ* **2020**, *368*, m441. [CrossRef] [PubMed]
35. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
36. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
37. Au, Q.; Herbinger, J.; Stachl, C.; Bischl, B.; Casalicchio, G. Grouped feature importance and combined features effect plot. *Data Min. Knowl. Discov.* **2022**, *36*, 1401–1450. [CrossRef]
38. Candès, E.; Fan, Y.; Janson, L.; Lv, J. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B* **2018**, *80*, 551–577. [CrossRef]
39. Watson, D.S.; Wright, M.N. Testing conditional independence in supervised learning algorithms. *Mach. Learn.* **2021**, *110*, 2107–2129. [CrossRef]
40. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.

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