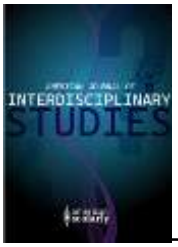


American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
Doi: 10.63125/fqs8ca04

USING MACHINE LEARNING TO ASSESS WORKPLACE ACCIDENT COMPENSATION CASES VIA HEALTH RECORDS

Rumaesa Mandana Rob[1]; Subaya Rahman Sabrin[2]

[1]. Master of Science in Information Technology (MsIT), Washington University of Science and Technology (WUST), Virginia, USA; Email: rumaesamandana.iub@gmail.com

[2]. Master of Science in Information Technology (MsIT), Washington University of Science and Technology (WUST), Virginia, USA; Email: ssabrin.student@wust.edu

ABSTRACT

This study employs a quantitative, cross-sectional, multi–case design to evaluate determinants of workplace accident compensation outcomes through harmonized electronic health records (EHRs), administrative claims, and a brief Likert-scale survey capturing process and communication constructs. The investigation encompasses multiple compensation jurisdictions and health systems, enabling comparative analysis and enhancing external validity. Closed adjudicated cases serve as the analytic unit, providing consistent labeling for four key outcomes, approval status, benefit magnitude, processing time, and appeal occurrence, each reflecting distinct stages of the adjudicative lifecycle. Data integration occurs under a privacy-preserving linkage protocol across three coordinated sources: structured EHR extracts containing diagnostic codes, procedures, medications, laboratory results, and encounter timestamps; claims system data containing lodgment and decision dates, indemnity and medical payments, attorney involvement, and prior claims indicators; and survey data capturing process-level constructs such as documentation completeness, communication quality, transparency, and adjudication clarity. Rigorous data quality assessments, reproducible phenotyping of clinical and administrative variables, and standardized reporting protocols ensure methodological transparency. Missingness is characterized and resolved through multiple imputation or indicator approaches, outliers are managed via clinically justified bounds, and all variables are timestamped relative to the index event to maintain temporal coherence and prevent information leakage. The analytical framework combines explanatory and predictive modeling to investigate the interplay between clinical burden, process efficiency, and adjudicative outcomes. Logistic regression models estimate approval and appeal likelihoods, generalized linear models with Gamma family and log link address skewed benefit magnitudes, and Cox proportional hazards models quantify time-to-decision processes. Model performance is evaluated using discrimination (AUC, precision–recall AUC, F1 score), calibration (Brier score, calibration slope), and equity diagnostics across subgroups defined by age, sex, industry, and language. Predictor domains include clinical severity, comorbidities, medication exposure, care intensity, process metrics, prior claims, and demographic–occupational context, complemented by the four Likert-based process indices. Reproducibility is ensured through version-controlled scripts, documented variable definitions, and auditable preprocessing pipelines that promote transparency and external validation. Collectively, the study advances a robust, equity-aware methodological framework for assessing and predicting compensation adjudication outcomes using EHR-linked data—bridging clinical analytics with administrative decision-making to enhance fairness, interpretability, and accountability in occupational health research.

KEYWORDS

Workplace Accident Compensation; Electronic Health Records; Administrative Claims; Quantitative Cross-Sectional

Citation:
Rob, R. M., & Sabrin, S. R. (2023). Using machine learning to assess workplace accident compensation cases via health records. *American Journal of Interdisciplinary Studies, 4*(2), 29–63. https://doi.org/10.63125/fqs8ca04

Received:
March 17, 2023

Revised:
April 20, 2023

Accepted:
May16, 2023

Published:
June 28, 2023

Copyright:

© 2023 by the author. This article is published under the license of American Scholarly Publishing Group Inc and is available for open

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

INTRODUCTION

Workplace accidents are commonly defined as unplanned, discrete events arising out of or in the course of employment that result in injury, illness, or death, with consequences that may be immediate or delayed and that frequently trigger statutory compensation processes under workers' compensation schemes (International Labour Organization definitions are widely used in policy and surveillance, but empirical burden estimates are largely derived from global epidemiology studies). Electronic health records (EHRs) refer to longitudinal, digital clinical repositories that include diagnoses, procedures, medications, laboratory values, and encounter timelines; when linked to claims and administrative adjudication files, they enable high-resolution measurement of injury mechanisms, severity, treatment, and recovery (Rajkomar et al., 2019). In quantitative research, Likert's five-point rating format (Strongly disagree→Strongly agree) is frequently applied to elicit process- or perception-based constructs such as documentation completeness or perceived fairness, which can be validated for internal consistency and construct coherence prior to analysis (Tavakol & Dennick, 2011).

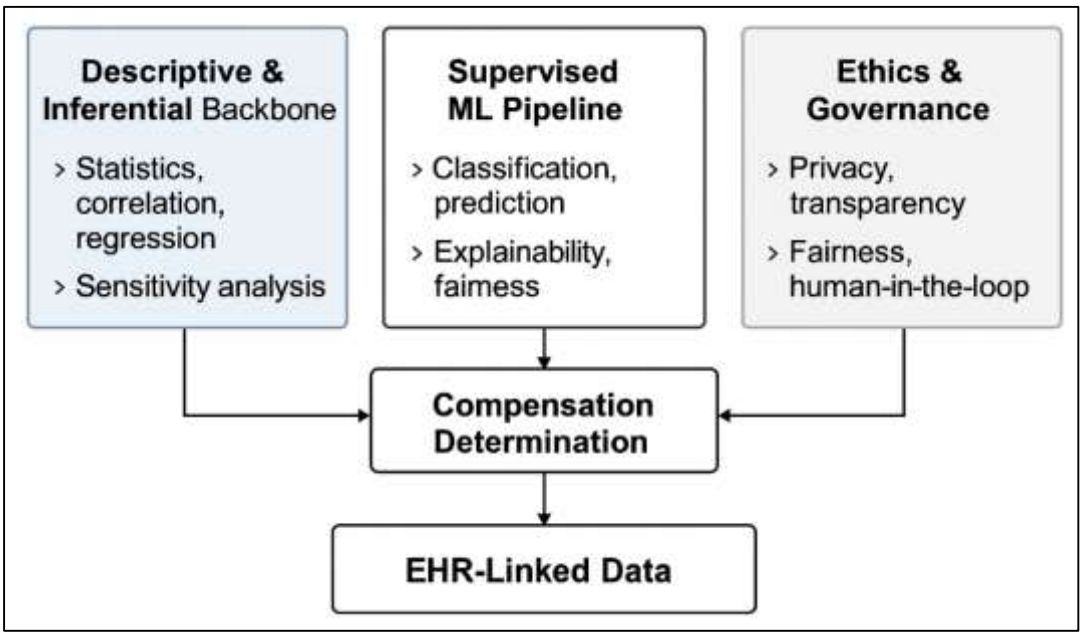
Figure 1: Worker's Compendation Claims Process

At a population level, the international significance of workplace injuries is underscored by the Global Burden of Disease consortium, which has reported millions of disability-adjusted life years (DALYs) attributable to occupational injuries annually and substantial cross-country heterogeneity by industry, sex, and age (Gebru et al., 2021). For payers, employers, and regulators, workers' compensation operates as a social insurance mechanism to finance medical care and wage replacement and to adjudicate eligibility and award magnitudes, yet empirical research shows considerable variability across jurisdictions and sectors in both approval decisions and benefit amounts (Smith et al., 2023). Against this backdrop, machine learning (ML) and modern regression modeling offer complementary lenses: regression for estimating effect sizes and testing hypotheses about specific determinants, and ML for optimizing predictive performance and uncovering

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

complex, nonlinear patterns in high-dimensional EHR–claims features (Breiman, 2001). Together, these tools support a transparent, data-driven approach to assessing compensation cases via health records while maintaining statistical interpretability and clinical relevance through calibration, validation, and human review (Collins et al., 2015).

Figure 2: HER-Linked Data



Empirical studies indicate that approval rates, processing times, and award magnitudes differ not only across jurisdictions but also across worker subgroups, even after accounting for injury severity and industry risk, highlighting potential inequities and process frictions (Charlson et al., 1987). Administrative datasets from state and national workers' compensation systems reveal patterns such as lower claim rates or lower benefit awards among certain racial or ethnic groups in comparable risk environments, raising concerns about surveillance limitations and adjudication consistency (Obermeyer et al., 2019). Research in adjacent health-system resource allocation shows that algorithmic tools can inadvertently encode structural inequities for example, when cost is used as a proxy for clinical need, resulting in systematically under-prioritizing patients with equal burden of illness but historically lower spending (O'Brien, 2007; Obermeyer et al., 2019). Although that study focused on care management rather than compensation per se, the lesson generalizes: predictors and labels must be chosen and audited to avoid reproducing disparities. From a methodological standpoint, the combination of classical regression to quantify adjusted associations (e.g., odds ratios for approval as a function of severity, documentation completeness, attorney involvement, and comorbidities) and ML to improve predictive accuracy and case triage provides a pragmatic strategy that keeps both fairness and explainability in view (Hosmer & Lemeshow, 1980). Public health surveillance work also documents sector-specific hazards (e.g., heat illness, construction, and manual trades), which translate into different injury profiles and compensation trajectories, necessitating models that incorporate industry codes and job tasks to avoid ecological confounding (Chen & Guestrin, 2016). Consequently, this study positions equity-aware modeling paired with subgroup calibration checks and transparent reporting as central to credible, practice-relevant assessment of compensation outcomes from EHR-linked data.

Rigorous assessment depends on accurate operationalization of outcomes and predictors. Compensation outcomes can be framed as approval (yes/no), benefit magnitude (continuous, often right-skewed), processing time (time-to-decision), and appeals (yes/no). EHR-derived predictors include diagnosis groupings (ICD injury chapters), procedure categories (e.g., imaging, surgery), medication classes, vital-sign abnormalities, and laboratory thresholds; claims-system predictors include time-to-filing, attorney involvement, and prior claims history. Clinical burden and baseline health status can be summarized with validated comorbidity indices such as the Charlson

Comorbidity Index (Mitchell et al., 2019) and injury severity may be proxied by measures like the Glasgow Coma Scale for head injury or triage acuity flags where applicable (Teasdale & Jennett, 1974). When constructing perception or process indices from Likert items (e.g., documentation completeness, communication quality, transparency, fairness), internal consistency should be appraised (Cronbach's α≥0.70) and dimensionality examined via exploratory/confirmatory factor analysis before creating composite scores (Peduzzi et al., 1996). On the informatics side, prior syntheses show that EHR data are heterogeneous and longitudinal, requiring careful handling of missingness, code mapping, and temporal alignment; still, they have repeatedly supported robust clinical prediction when modeled with both regression and ML (Shickel et al., 2018; Wong et al., 2023). Calibration and external validity are critical: models that appear accurate in internal testing can fail when transported across sites with different patient mixes, coding practices, or compensation rules, hence the need for explicit calibration assessment and transparent reporting (Snell et al., 2023; Wrona, 2006). This study's variable map therefore spans clinical, process, demographic, and employment dimensions, enabling both explanatory analyses and predictive pipelines while retaining a governance framework for measurement quality and privacy.

A quantitative, cross-sectional, multi-case design is suitable when the goal is to evaluate associations and build/benchmark predictive models across diverse organizational contexts within a defined time window. Cross-sectional snapshots that include all closed cases within a period allow consistent labeling of outcomes (e.g., final approval and benefit amounts) and standardization of exposure windows (e.g., time-to-first treatment). Multi-case sampling across sites and jurisdictions enhances external validity by capturing heterogeneity in practices and regulations while enabling cluster-robust standard errors and fixed-effects adjustments in regression models. Within this framework, survey-based Likert measures collected from clinicians or adjudicators can quantify process quality constructs documentation completeness, communication, and transparency that are otherwise difficult to infer from structured fields alone; reliability and construct validity procedures reduce measurement error and support hypothesis testing (Teasdale & Jennett, 1974). To safeguard reproducibility and clinical credibility, reporting should adhere to consensus guidelines for prediction model studies, including the TRIPOD statement for regression- and ML-based models and its extensions for systematic reviews (Manning & Mullahy, 2001). Additionally, ML deployment in clinical workflows requires design choices that surface predictions and explanations at the right place and time within EHR systems, with evidence accumulating on best practices for safe integration and evaluation (Wang et al., 2024; Zou & Hastie, 2005). Because compensation adjudication intersects clinical, legal, and administrative domains, the study's design emphasizes both model performance and interpretability (e.g., marginal effects in regression; feature importance and SHAP-style local explanations in ML), while instituting subgroup analyses and calibration checks to monitor consistency across age, sex, and industry strata (Calster et al., 2019). Together, these design elements support rigorous, transparent evaluation of determinants and predictive models using EHR-linked compensation data.

The inferential backbone of the study comprises descriptive statistics, correlation analyses, and multivariable regression modeling. Descriptive summaries (means/SDs or medians/IQRs for continuous variables; counts/percentages for categorical variables) stratified by approval status provide initial contrasts. Correlation matrices (Pearson or Spearman as appropriate) among continuous predictors and Likert indices help identify collinearity before modeling. For the binary approval outcome, logistic regression estimates adjusted odds ratios with 95% confidence intervals; goodness-of-fit and calibration are appraised via the Hosmer–Lemeshow framework and calibration plots, supplemented by cluster-robust standard errors at the site level. For the highly skewed benefit magnitude outcome, generalized linear models with Gamma family and log link, or log-OLS with smearing retransformation, provide consistent estimation under heteroskedasticity (Bonauto et al., 2006). Processing time can be analyzed via accelerated failure time models or Cox proportional hazards models if time-to-event assumptions are met. To preserve interpretability and avoid overfitting, variable selection can combine domain knowledge with penalized methods such as elastic net for shrinkage (Zou & Hastie, 2005), while maintaining minimum "events-per-variable" thresholds derived from simulation and empirical guidance (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996). Multicollinearity can be monitored using variance inflation factors, with careful interpretation of "rules of thumb" (Mitchell et al., 2019). Across models, partial R^2 and marginal effects facilitate substantive interpretation. Finally, sensitivity analyses alternative missingness handling,

exclusion of extreme severities, and jurisdiction fixed effects test robustness, while explicit calibration assessment addresses the well-documented Achilles' heel of predictive analytics (Calster et al., 2019).

To complement explanatory regression, a supervised ML pipeline will target (a) approval classification and (b) high-cost prediction (e.g., top decile of benefit). Baselines include regularized logistic regression and elastic net; tree-based ensembles (random forests; gradient boosting) offer capacity for nonlinear interactions among EHR-derived features and process indices (Smith et al., 2023). Model development uses a stratified train/validation/test split with k-fold cross-validation and hyperparameter tuning; performance is summarized with area under the ROC curve (AUC), precision–recall AUC (for class imbalance), F1 score, and calibration metrics (e.g., Brier score and reliability curves). Transparent reporting per TRIPOD now widely cited in prediction research supports reproducibility and appraisal of risk of bias, while newer extensions address synthesis and AI-specific reporting details (Collins et al., 2015; Snell et al., 2023). Clinical ML experiences further argue for integration considerations within EHRs to ensure that predictions are delivered at the point of need and paired with human review (Rajkomar et al., 2019). Explainability is addressed with global feature importance and local explanation techniques (e.g., SHAP), paired with partial dependence or individual conditional expectation plots to visualize marginal effects; these tools aid adjudicators in understanding drivers of predictions and in aligning model outputs with statutory criteria. Crucially, subgroup performance and calibration will be evaluated by age, sex, and industry, given evidence that algorithms trained on administrative or cost-based proxies can differentially rank patients across demographic groups (Obermeyer et al., 2019). Collectively, this pipeline is designed to maximize predictive utility while preserving interpretability, calibration, and fairness key prerequisites for any decision-support augmentation of compensation adjudication.

Because compensation determinations carry material and legal consequences, the modeling enterprise must embed ethics and governance from the outset. First, data protection requires de-identification, secure linkage of EHR and claims data, and auditable access controls consistent with health-information privacy norms. Second, model development should be documented with "model cards" and dataset documentation to support accountability and stakeholder scrutiny (Mitchell et al., 2019). Third, fairness requires explicit subgroup evaluation and, when needed, mitigation (e.g., reweighting, threshold adjustments) to prevent disparate error rates or calibration drift across protected groups, as cautioned by high-profile evidence of bias propagation when cost or utilization proxies are substituted for clinical need (Obermeyer et al., 2019). Fourth, transparent reporting standards (TRIPOD) and their newer extensions for prediction model syntheses provide a framework to disclose dataset composition, missingness, modeling choices, and validation results (Collins et al., 2015). Finally, human-in-the-loop procedures where adjudicators use model outputs as decision aids rather than replacements align with emerging clinical ML guidance emphasizing workflow integration, continuous monitoring, and post-deployment calibration checks (Wong et al., 2023). In combination, these practices position the present study to contribute credible evidence on determinants and predictive assessment of compensation outcomes using EHR-linked data, while honoring equity and due-process considerations fundamental to workers' compensation systems.

The objective of this study is to develop a rigorous, transparent, and reproducible framework for assessing workplace accident compensation cases using linked electronic health records and administrative claims data, complemented by standardized Likert-based measures of process quality and perceived fairness. Specifically, the study aims to: (1) operationalize a comprehensive variable map that captures clinical severity, comorbidity profiles, diagnostic and procedural information, timelines of care and filing, prior claims history, attorney involvement, and worker and industry characteristics; (2) construct and validate internally consistent composite indices from a five–point Likert instrument to quantify documentation completeness, communication quality, transparency, and adjudication clarity; (3) describe the study population and included cases with detailed summary statistics stratified by compensation outcomes to establish baseline distributions and identify crude contrasts; (4) estimate adjusted associations between prespecified predictors and four primary outcomes approval status, benefit magnitude, processing time, and appeal occurrence using appropriate regression families with cluster-robust standard errors, interactions for moderation, and sensitivity analyses for missingness and model specification; (5) build and evaluate supervised machine learning models for approval classification and high-cost prediction, using stratified data splits, cross-validation, and hyperparameter tuning, and report discrimination,

calibration, and error characteristics alongside interpretable summaries of feature influence; (6) compare classical regression and machine learning performance on common test sets to determine added predictive value beyond parsimonious, theory-driven models; (7) examine consistency of associations and predictions across demographic and industry subgroups, quantify calibration and error parity, and document any performance differentials; (8) formalize a measurement and governance protocol that includes de-identification procedures, auditable data linkage, code versioning, analysis scripts, and structured model documentation to enable replication; and (9) synthesize the empirical findings into a concise set of evidence-based decision aids such as marginal effect summaries, risk score thresholds aligned to defined performance targets, and construct scores derived from the survey that can be incorporated into standardized case review workflows. Collectively, these objectives establish a coherent plan to quantify determinants, benchmark predictive performance, ensure reliability and validity of key measures, and deliver an interpretable, ethics-aware assessment framework for evaluating compensation cases via health records in a cross-sectional, multi–case setting.

LITERATURE REVIEW

The literature on workplace accident compensation at the intersection of occupational health, health informatics, and decision science spans three strands that converge on a common challenge: how to turn heterogeneous clinical and administrative information into reliable assessments of claim outcomes. First, occupational health research establishes the burden and distribution of workplace injuries across industries and job tasks, identifies clinical and demographic determinants of severity, and documents variation in claim approval, benefit magnitude, and time-to-decision. These studies emphasize the roles of injury mechanism, comorbidity, documentation completeness, and attorney involvement, while also noting jurisdictional rules and organizational practices that shape adjudication. Second, health informatics research demonstrates that electronic health records, when methodically curated and linked to claims, support high-resolution measurement of diagnoses, procedures, medications, and timelines of care. This stream contributes methods for code mapping, episode construction, handling missingness, and summarizing clinical burden with indices such as comorbidity or acuity scores, laying a measurement foundation for both explanatory and predictive modeling. Third, decision science and predictive analytics introduce statistical and machine learning techniques to estimate associations, classify outcomes, and forecast costs, paired with calibration, validation, and explainability procedures that determine whether models generalize across settings and subgroups. Within this landscape, survey-based instruments using Likert scales enrich structured data by capturing process constructs documentation quality, communication, and transparency that are central to adjudication yet underrepresented in routine EHR fields. Together, these strands reveal consistent themes: the need for clearly defined outcomes and predictors; rigorous preprocessing and linkage of clinical and claims data; analytic strategies that combine regression for inference with machine learning for pattern discovery; and reporting practices that foreground calibration, reliability, and subgroup consistency. At the same time, the literature underscores the practical constraints of cross-sectional designs, the importance of sample size relative to model complexity, and the value of standardized codebooks and model documentation for reproducibility. This review positions the present study within that evidence base, synthesizing key determinants, measurement practices, and analytic frameworks to support a transparent, interpretable, and ethically grounded approach to assessing workplace accident compensation cases via health records.

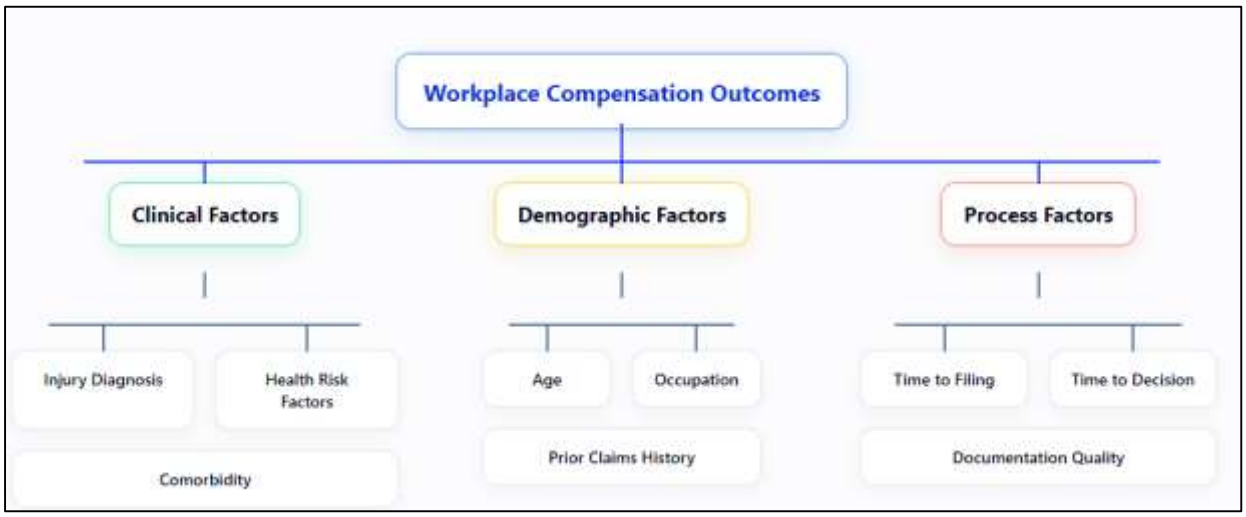
Determinants of Workplace Compensation Outcomes

At the core of workers' compensation adjudication is a set of clinical, demographic, and process factors that shape whether a claim is approved, how long it remains active, and the ultimate magnitude of wage-replacement and medical benefits. Evidence drawn from population-level claims shows that sustained return-to-work (RTW) and related outcomes depend on measurable attributes such as injury diagnosis, prior health status, occupation, age, and the structure of wage-replacement rules, indicating that adjudication is responsive to both clinical severity and worker context. In a large study of more than 59,000 claims, predictors of sustained RTW included injury type, industry, age, and prior claims history, illustrating how case-mix and employment characteristics intersect with administrative decisions to influence disability trajectories (Berecki-Gisolf et al., 2012). Beyond canonical injury metrics, worker health profiles captured via health-risk factors such as smoking, obesity, stress, and physical inactivity have been linked to both the occurrence of

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

compensable injuries and claim costs, underscoring that baseline health shapes hazard exposure, recovery, and treatment intensity recorded in health records and, by extension, compensation outcomes (Schwatka et al., 2017). Together, these findings motivate variable maps that include not only diagnoses and procedures but also comorbidity and risk-factor summaries, occupation and industry codes, and prior utilization, because these features jointly determine the likelihood of approval, the speed of case resolution, and the size of indemnity and medical payments. They also justify modeling strategies that can parse overlapping determinants while retaining interpretability for adjudicators who must apply statutory criteria consistently across heterogeneous cases (Berecki-Gisolf et al., 2012).

Figure 2: Determinants of Workplace Compensation Outcomes Framework



Process factors embedded in the life cycle of a claim further modulate outcomes by shaping the timing of decisions and the duration of compensated time loss. Population-level analyses demonstrate that claim processing intervals such as time from lodgment to decision and from decision to first wage-payment are associated with longer disability duration, indicating that administrative delays are not merely bureaucratic artifacts but measurable determinants of worker recovery and financial exposure within compensation systems. In state-wide data, longer lodgment, decision, and total processing times were linked to higher hazards of extended disability, suggesting that timely adjudication and payment can reduce the period workers remain off work and thereby moderate downstream costs and appeals (Gray et al., 2019). Administrative expectations recorded early in the case also carry predictive information: when claims managers document anticipated time to RTW within the first month, those expectations correlate with the actual duration of compensated disability, implying that early assessments capture meaningful signals about case complexity, documentation completeness, workplace accommodation, and clinical trajectory (Young et al., 2016). These observations reinforce the importance of modeling temporal features available in electronic health records and claims time to first treatment, time to filing, and decision intervals alongside clinical and demographic predictors, because delays can propagate into prolonged disability and greater indemnity exposure. They also point to a practical role for predictive analytics as triage aids: by identifying cases at risk of extended processing or prolonged time loss, administrators and clinicians can prioritize documentation audits, communication, and coordination to keep cases on a timely path, within statutory frameworks (Young et al., 2016).

Social and communication dimensions intersect with the determinants above by influencing documentation quality, access to care, and comprehension of the claims process. Language proficiency, for instance, has been shown to relate to patterns of acceptance, time loss, and cost among claimants with common musculoskeletal conditions, highlighting how limited English proficiency can introduce barriers to timely filing, clear clinical narratives, and effective navigation of procedures that require precise forms and consistent terminology (Bonauto et al., 2010). In practice, these barriers manifest in the health record and administrative file as missing or inconsistent data, longer intervals between events, and greater reliance on third-party assistance, which can in

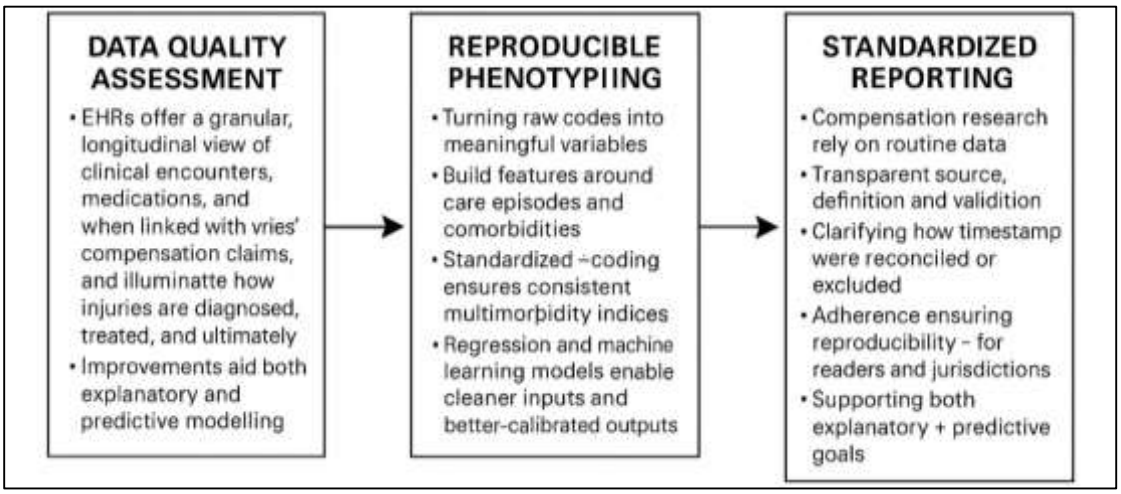
American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

turn affect approval odds, processing time, and award magnitude. Incorporating structured measures of communication quality and documentation completeness obtained via short Likert-scale instruments offers one path to capturing these dimensions systematically for use in explanatory and predictive models. When combined with robust clinical features, temporal markers, and employment context, these social and communication constructs provide a more complete representation of the determinants that shape compensation outcomes. They also underscore the need for reliability-tested indices that can be interpreted by adjudicators and claims professionals, ensuring that the analytics remain faithful to statutory criteria while reflecting the practical realities of case handling across diverse worker populations. In sum, contemporary evidence supports a determinants framework that integrates clinical burden, temporal process metrics, and communication context to explain variation in approval, disability duration, and payment levels across compensation systems (Gray et al., 2019; Young et al., 2016).

Electronic Health Records and Claims Analytics in Occupational Health

Electronic health records (EHRs) offer a granular, longitudinal view of clinical encounters, medications, procedures, and care timelines that, when linked with workers' compensation claims, can illuminate how injuries are diagnosed, treated, and ultimately adjudicated. For occupational health research, a persistent barrier to reusing these routinely collected data is variable data quality across sites and systems issues of completeness, correctness, timeliness, and concordance that can bias estimates or weaken prediction if left unmeasured. A pragmatic starting point is to treat data quality assessment as a first-order step in any analytic workflow, specifying which dimensions of quality are relevant to each variable (e.g., whether time-to-first-treatment requires event timestamp completeness, or whether diagnostic grouping requires coding consistency) and then selecting appropriate assessment methods before modeling. Doing so recognizes that compensation-related outcomes approval, benefit magnitude, processing time, appeals are not simply clinical endpoints; they emerge from administrative processes whose traces in EHRs and claims systems can be irregular (Weiskopf & Weng, 2013). Conceptual reviews of EHR data quality provide taxonomies and practical checks that help researchers align variables and study questions, encouraging explicit documentation of how each feature was defined, what portion of records met minimal quality thresholds, and how missingness was handled. In the compensation context, this means auditing the fidelity of injury codes, verifying key timestamps used to derive process intervals, and clarifying when narrative notes must be transformed (e.g., via structured abstraction) to quantify documentation completeness. Such preparatory work improves both explanatory modeling (e.g., regression) and predictive modeling (e.g., machine learning), because it reduces avoidable bias and clarifies the provenance of the analytic dataset (Weiskopf & Weng, 2013).

Figure 3: Electronic Health Records and Claims Analysis



Beyond quality screening, EHR–claims analytics in occupational health depend on reproducible "phenotyping" of injuries, comorbidities, and process markers turning raw codes and timestamps into clinically meaningful variables that can support association tests and risk stratification (Jensen et al., 2012). Phenotyping frameworks emphasize that raw EHR entries reflect both patient physiology and

care processes (what was measured, when, and why), so constructing valid features requires attention to ordering, temporality, and recording artifacts. In practice, this means defining episodes of care around the index injury, aligning encounter windows to filing dates, and deriving interpretable predictors such as imaging performed in the first 24–48 hours, opioid exposure in the acute phase, or referral patterns to specialty care, each with explicit rules. Critically, many occupational cohorts also require principled comorbidity adjustment because background health status shapes recovery, utilization, and claim trajectories; standardized coding algorithms allow Charlson or related indices to be computed consistently across ICD-9-CM/ICD-10 eras, ensuring that models control for multimorbidity in ways comparable across systems. When such phenotypes and indices are built with reproducible code and clear definitions, regression models can estimate adjusted associations with tighter interpretability, while machine learning pipelines can ingest richer, cleaner feature sets and yield better-calibrated predictions (Hripcsak & Albers, 2013; Rezaul, 2021). For claims administrators and clinicians, this translates into outputs odds ratios, partial effects, or risk scores whose inputs are auditable and whose construction rules can be shared across sites, sustaining trust and portability (Danish & MZafor, 2022; Hripcsak & Albers, 2013).

Finally, because compensation research relies on observational, routinely collected data, credible analytics require transparent reporting that discloses data sources, linkage procedures, variable definitions, and validation steps in a structured, journal-ready format. Reporting guidance tailored to routinely collected health data underscores items that are especially salient to EHR–claims studies: describing data provenance and any transformations; detailing inclusion and exclusion criteria tied to administrative states (e.g., "closed" or "adjudicated" claims); explaining how missing timestamps or inconsistent codes were resolved; and presenting sensitivity analyses that show the stability of results to alternative definitions or linkage rules (Jahid, 2022; Langan et al., 2018; Quan et al., 2005). For occupational health specifically, such guidance helps readers understand whether cross-site differences reflect true practice variation or artifacts of coding and data capture, and whether models have been fairly calibrated across demographic and industry subgroups. Adherence to these reporting standards also benefits downstream reproducibility: codebooks, phenotype definitions, and analysis scripts can be registered or shared, enabling other investigators and jurisdictions to repeat or extend findings with minimal reinvention (Danish & MKamrul, 2022; Langan et al., 2018; Quan et al., 2005). This culture of transparency supports both explanatory goals (e.g., testing hypotheses about determinants such as time-to-care or documentation completeness) and predictive goals (e.g., triaging cases at risk for extended processing) by making assumptions and data decisions legible to adjudicators, clinicians, and policymakers who must interpret and act on the results. In sum, for EHR–claims analytics to advance practice in workplace compensation, studies should pair rigorous data quality assessments with reproducible phenotyping and standardized reporting, thereby aligning methodological robustness with the operational realities of case review (Langan et al., 2018).

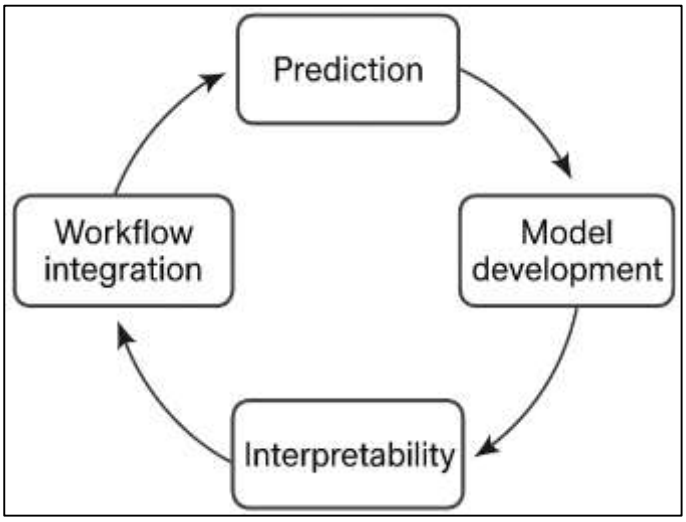
Machine Learning in Healthcare Decision Support

Machine learning (ML) in healthcare is best understood as a toolbox for converting routinely collected data structured codes, laboratory values, medications, and care timelines into predictions that can surface risks, prioritize review, and standardize assessments. In contrast to hand-crafted scoring rules, modern ML can flexibly capture nonlinearities and interactions among clinical and process variables relevant to adjudication (e.g., time-to-first-treatment, injury code patterns, prior claims). Within health systems, this promise has been framed as part of a broader shift toward data-intensive decision support, where algorithms complement clinician and administrator judgment by consistently highlighting high-risk cases and summarizing complex histories for timely action (Beam & Kohane, 2018; Ismail, 2022). For compensation research, the value proposition is twofold: first, ML can raise predictive sensitivity for outcomes like denial risk or prolonged processing time without unduly sacrificing calibration; second, it can integrate heterogeneous predictors from comorbidity indices and medication exposures to documentation signals to reflect the multifactorial nature of adjudication. Yet performance alone is not sufficient; decision-support tools must communicate uncertainty and avoid overfitting to site-specific idiosyncrasies. Emerging translational perspectives emphasize building models that are not only accurate but also human-aligned: they should be auditable, transportable across organizations, and designed so that predictions are easy to inspect within existing review workflows (Topol, 2019). In this sense, ML becomes a means to deliver standardized, reproducible triage cues to claims professionals, while preserving the adjudicator's

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

role in applying statutory criteria to fully documented cases (Beam & Kohane, 2018; Hossen & Atiqur, 2022).

Figure 4: Cycle of Machine Learning in Healthcare Decision Support



From a methods standpoint, rigorous model development begins with careful data partitioning and sample-size planning, followed by honest internal validation and clear reporting of discrimination and calibration. Observational datasets typical of EHR–claims linkages often contain thousands of cases but relatively few outcome events for rarer endpoints; ignoring this imbalance can inflate apparent performance. Contemporary guidance on minimum sample size for multivariable prediction stresses controlling model optimism by ensuring enough events per candidate parameter, constraining model complexity, and using bootstrap or cross-validation to quantify overfitting (Kamrul & Omar, 2022; Riley et al., 2020). Metric choice matters as well: when outcomes are imbalanced (e.g., appeals or high-cost tail), the area under the precision–recall curve (PR-AUC) provides a more faithful picture of utility than accuracy or ROC AUC, because it focuses attention on positive predictive value at relevant recall levels (Razia, 2022; Saito & Rehmsmeier, 2015). For binary endpoints central to compensation (approval, appeal), reporting threshold-based measures (precision, recall, F1) alongside calibration plots helps stakeholders select operating points that balance false negatives and false positives in line with policy tolerances; for continuous targets (benefit magnitude), error and calibration summaries (e.g., MAE with calibration-in-the-large) should be provided on appropriately transformed scales. Throughout, reproducible preprocessing handling missingness, defining episodes, aligning exposure windows is essential to preserve transportability. Models that pass internal validation should, where feasible, be stress-tested across sites to detect calibration drift before any consideration of operational use (Ribeiro et al., 2016; Sadia, 2022; Saito & Rehmsmeier, 2015).

Interpretability is pivotal for gaining trust among clinicians and compensation professionals who must explain and defend decisions. Post-hoc explanation techniques such as local surrogate models can provide case-level rationales by approximating the model's behavior in the neighborhood of a specific prediction, surfacing which features drove a denial-risk alert or a prolonged-processing forecast (Danish, 2023; Topol, 2019). However, explanations are most useful when paired with inherently more intelligible modeling strategies (e.g., monotonic generalized additive models or sparsity-promoting linear models) so that global behavior remains comprehensible and stable; the goal is not merely to "open the black box," but to ensure that the model's structure aligns with domain knowledge and statutory logic. At the interface, visualization of partial effects and interaction patterns can help reviewers sanity-check learned relationships (for example, that longer time-to-care monotonically raises risk within plausible bounds) and detect artifacts that warrant data or feature engineering refinement. Translational commentaries argue that successful clinical decision support blends three elements: validated predictive performance, interpretable evidence of "why" a prediction was made, and careful workflow integration that reduces cognitive load rather than adding alerts (Arif Uz & Elmoon, 2023; Riley et al., 2020). For the present research, adopting a layered interpretability strategy global summaries for governance, local rationales for case review supports

an ethics-aware deployment posture in which models inform but do not replace human adjudication (Hossain et al., 2023; Ribeiro et al., 2016; Topol, 2019).

Fairness in Algorithmic Decision-Making

A credible framework for assessing workplace accident compensation with electronic health records must embed fairness and transparency from the outset, because predictive tools can differentially advantage or disadvantage subgroups even when overall accuracy appears high. Foundational work in algorithmic fairness shows that widely used group-fairness criteria such as equal error rates, calibration across groups, and predictive parity cannot, in general, be satisfied simultaneously when outcome base rates differ across groups. This "impossibility" result matters for compensation adjudication, where injury types, occupations, and prior health burdens can legitimately vary by worker subgroup; choosing one fairness target may necessarily relax another, so governance must make value trade-offs explicit rather than implicit (Hasan, 2023; Selbst et al., 2019). In parallel, scholarship on disparate impact demonstrates how seemingly neutral models can produce unequal error distributions that translate into real-world harms when decisions are thresholded, audited, or operationalized at scale. Taken together, these insights push compensation analytics beyond a purely technical exercise toward a policy-aware design in which fairness targets are selected, justified, and monitored in light of statutory aims (e.g., timely, equitable access to benefits) and practical constraints (e.g., class imbalance, noisy labels). In this section, we adopt those insights to shape a governance approach that pairs statistical reporting (discrimination and calibration) with fairness reporting (error parity and calibration-in-the-large by subgroup), recognizing that attaining one desideratum may preclude others and that such conflicts must be surfaced, documented, and overseen rather than assumed away (Chouldechova, 2017; Shoeb & Reduanul, 2023; Raji & Buolamwini, 2019).

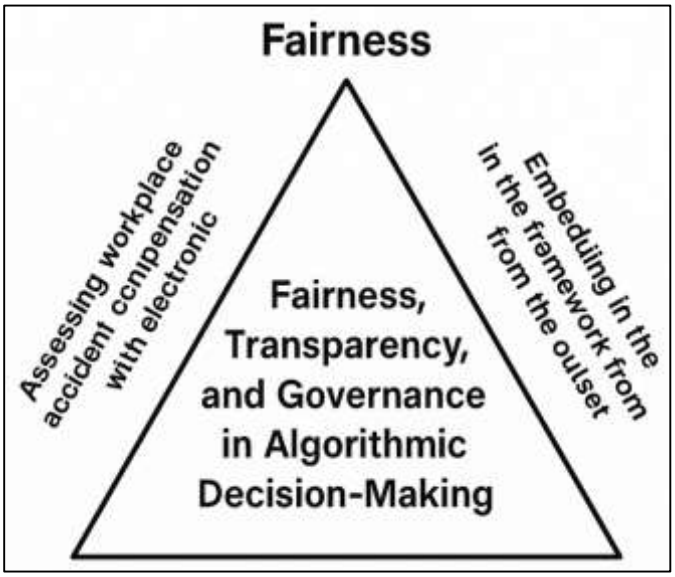
Fairness is also inseparable from context: models are deployed inside sociotechnical systems comprising institutions, policies, documentation practices, and human workflows. A sociotechnical perspective cautions against "abstraction traps" in which designers treat fairness as a property of the model alone, ignoring how data are generated (e.g., which injuries get coded or contested), how documentation is produced (e.g., language proficiency shaping narrative completeness), and how predictions are interpreted (e.g., local thresholds or appeal processes) (Kleinberg et al., 2017; Mubashir & Jahid, 2023; Selbst et al., 2019). In workers' compensation, this means that fairness cannot be guaranteed by adding a constraint in training; it requires an end-to-end view of case intake, evidence gathering, communication, and adjudication. Governance mechanisms should therefore include: (i) dataset documentation that explains provenance, sampling, and known limitations; (ii) pre-specified subgroup audits of calibration and error rates by age, sex, industry, and language preference; (iii) procedures to monitor drift as coding practices, clinical pathways, or filing behaviors evolve; and (iv) channels for contestability, so that claimants and adjudicators can interrogate predictions with accessible rationales (Razia, 2023). This approach reframes fairness as an organizational capability regular audits, clear escalation paths, and transparent documentation rather than a one-time technical fix. By situating models within the realities of case handling, appeal timelines, and documentation ecology, sociotechnical governance helps ensure that improvements in discrimination or speed do not come at the cost of legitimacy, appealability, or due process (Kleinberg et al., 2017; Reduanul, 2023; Selbst et al., 2019).

Finally, transparency requires both external accountability and internal reproducibility. Public-facing algorithmic auditing has demonstrated that systematic, evidence-based disclosures about error disparities can shift commercial behavior and catalyze corrective action an instructive precedent for compensation systems that must account to workers, employers, and regulators. Internally, reporting standards developed for artificial-intelligence interventions in healthcare trials highlight what robust documentation looks like: clear descriptions of data sources and eligibility, prespecified outcomes, handling of missingness, and precise accounts of model versioning, human oversight, and failure modes. While compensation analytics are observational rather than randomized, the same spirit applies: declare in advance how fairness will be assessed, how thresholds will be chosen, how explanations will be presented to decision makers, and how updates will be governed. Concretely, this entails model cards that include subgroup performance; operating-point justifications tied to policy goals; and audit logs that record when and how predictions influenced case review. Embedding such practices makes the predictive pipeline inspectable, reproducible, and aligned with the system's normative commitments (Sadia, 2023). In sum, fairness, transparency, and

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

governance are not add-ons to an analytical pipeline but coequal design objectives that shape data curation, modeling choices, evaluation, and deployment thereby supporting equitable, defensible use of EHR-linked analytics in workplace accident compensation (Liu et al., 2020; Raji & Buolamwini, 2019).

Figure 5: Triangle Framework of Fairnessin Algorithmic Decision-Making



METHOD

This study adopts a quantitative, cross-sectional, multi–case design to evaluate determinants of workplace accident compensation outcomes using linked electronic health records (EHRs), administrative claims, and a brief Likert-scale instrument capturing process and communication constructs. The target population comprises closed compensation cases adjudicated within a predefined observation window across multiple organizations or jurisdictions, enabling consistent outcome labeling (approval status, benefit magnitude, processing time, and appeal occurrence) and comparative analysis across sites. Data integration proceeds in three layers. First, structured EHR extracts provide diagnoses, procedures, medications, laboratory results, and encounter timestamps; these are curated into clinically meaningful phenotypes (e.g., injury groupings, comorbidity indices, markers of acute care intensity). Second, claims data contribute filing and decision timestamps, indemnity and medical payments, prior claims history, attorney involvement, and administrative statuses required to operationalize key process intervals. Third, a short, reliability-tested survey administered to involved clinicians or claims officers yields four 5-point Likert indices documentation completeness, communication quality, transparency, and adjudication clarity designed to quantify process features underrepresented in routine data. Case inclusion, exclusion, and deduplication rules are prespecified, with deterministic or probabilistic linkage used to join EHR and claims records under privacy-preserving protocols. Following data cleaning, missingness is profiled and addressed using appropriate strategies (e.g., indicator methods for sparse categorical fields; multiple imputation for continuous variables where assumptions permit), and outliers are screened according to defensible clinical or administrative bounds. The statistical analysis plan combines descriptive summaries and group contrasts with correlation matrices to assess collinearity, followed by multivariable regression tailored to each outcome: logistic models for approval and appeal, generalized linear models for skewed benefit amounts, and time-to-event models for processing duration. Prespecified interaction terms test moderation (e.g., severity × documentation completeness), and cluster-robust standard errors account for site-level heterogeneity. In parallel, supervised machine learning pipelines (regularized linear models and tree-based ensembles) are trained on harmonized feature sets using stratified splits and cross-validation, with discrimination, calibration, and error characteristics reported on hold-out tests. Subgroup diagnostics evaluate performance and calibration across age, sex, industry, and language groups. Reproducibility is supported through a version-controlled workflow encompassing phenotype definitions, codebooks,

Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

preprocessing scripts, model specifications, and analysis notebooks, ensuring that all results are auditable and portable across participating sites.

Design: Quantitative

This study employs a quantitative, cross-sectional, multi–case design to assess determinants of workplace accident compensation outcomes using harmonized electronic health records (EHRs), administrative claims, and a brief Likert-scale survey capturing process and communication constructs. The cross-sectional frame is defined as all compensation cases that reached a closed adjudicative state within a prespecified observation window, ensuring consistent outcome labeling for four endpoints: approval status, benefit magnitude, processing time, and appeal occurrence. A multi–case structure comprising multiple organizations, sites, or jurisdictions intentionally samples heterogeneity in practice patterns, documentation cultures, and statutory rules; this enables comparative inference and enhances external validity while permitting cluster-robust estimation and site controls in modeling. The unit of analysis is the index claim associated with a qualifying workplace accident; where individuals have multiple incidents in the window, deterministic rules prioritize the first closed case or the most clinically severe event to avoid correlated outcomes. The design integrates three data layers under a privacy-preserving linkage protocol: (i) structured EHR extracts (diagnoses, procedures, medications, labs, encounters, timestamps) curated into clinically interpretable phenotypes and comorbidity indices; (ii) claims system fields (lodgment and decision timestamps, indemnity and medical payments, prior claims, attorney involvement, administrative statuses); and (iii) a short survey administered to clinicians or claims officers, yielding four five-point indices documentation completeness, communication quality, transparency, and adjudication clarity validated for internal consistency before analysis. Inclusion criteria encompass adult workers with a confirmed work-related injury, complete identifiers enabling EHR–claims linkage, and an adjudicated claim within the window; exclusions address non-occupational injuries, open or pending cases, and records lacking essential timestamps or outcome labels. The protocol prespecifies data cleaning, missingness handling, and outlier rules; defines site-level governance for access, audit, and de-identification; and registers primary and secondary analyses to minimize analytic flexibility. This design supports both explanatory regression (hypothesis testing with interpretable effects) and predictive modeling (supervised learning for triage and risk stratification), while embedding subgroup diagnostics to evaluate consistency across age, sex, industry, and language groups.

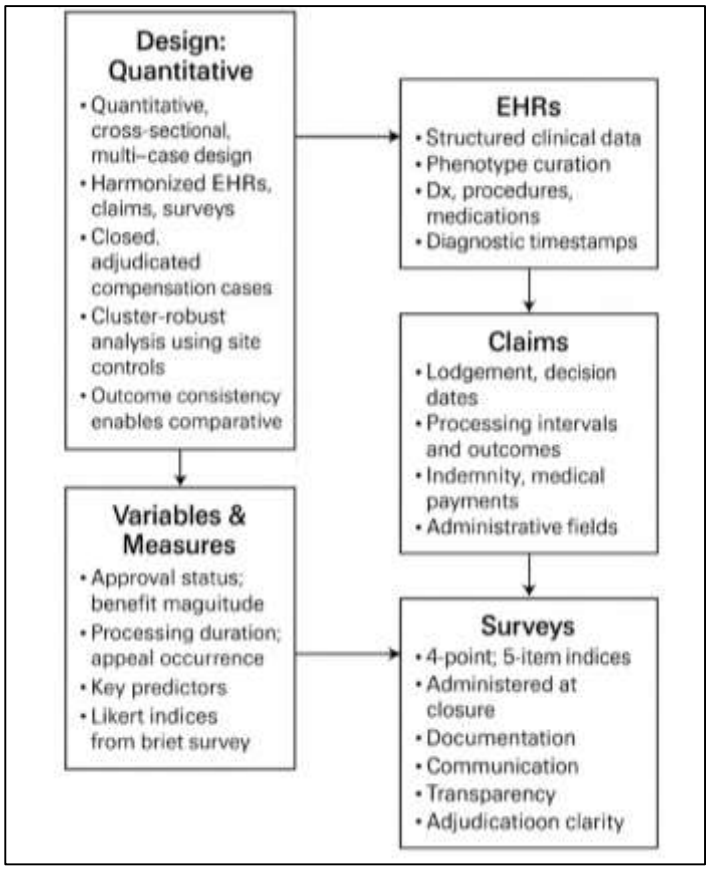
Cases, Sampling, and Setting

The study setting comprises multiple workers' compensation jurisdictions and affiliated health systems that maintain interoperable, queryable electronic health records (EHRs) and administrative claims platforms. Each participating site designates a data steward and an adjudication liaison to ensure that clinical, administrative, and legal variables are consistently interpreted. The analytic frame is a census of *closed* compensation cases reaching a final adjudicative state within the observation window (e.g., January 1, 20XX to December 31, 20YY). A "closed" case is operationalized as one with a recorded decision code and no pending appeal at data freeze; appealed cases are included when an appeal disposition is recorded within the window, allowing appeal occurrence and timing to be analyzed as outcomes. The *index event* for each case is the earliest employer-reported incident or, when unavailable, the first clinical encounter coded as work-related within ±7 days of the incident report. Data sources include structured EHR tables (diagnoses, procedures, medications, labs, encounter timestamps), claims tables (lodgment and decision dates, indemnity and medical payments, attorney involvement, prior claims indicators), and a short survey administered to treating clinicians or claims officers to capture four five-point indices: documentation completeness, communication quality, transparency, and adjudication clarity. Record linkage proceeds under a privacy-preserving protocol: deterministic joins use encrypted person identifiers, date of birth, and employer code; when a deterministic key is missing, a probabilistic linkage is attempted using encrypted quasi-identifiers with clerical review by the site steward. A uniform data dictionary and extract-transform-load (ETL) specification are distributed prior to onboarding; sites deliver de-identified extracts to a secure analysis enclave. Quality gates check field completeness, temporal coherence (e.g., filing before decision; treatment before decision), and code-set validity prior to lock. The *unit of analysis* is the index claim; when a worker has multiple closed claims in-window, deterministic rules select the most clinically severe or, if severities tie, the earliest-closed case to avoid within-person correlation in primary analyses.

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

Sampling is designed as an inclusive census of all eligible, closed claims within the study window to maximize precision for subgroup analyses and to preserve the natural prevalence of outcomes (approval, appeal, prolonged processing, high-cost tail). Sites with very small volumes are retained to reflect real-world heterogeneity; cluster-robust variance estimation and site fixed effects accommodate this structure in regression models. Anticipated sample size is guided by outcome frequencies provided during site scoping; for binary endpoints (approval, appeal), events-per-parameter thresholds inform the number of prespecified covariates and interactions, while continuous endpoints (benefit magnitude) target precision benchmarks expressed as maximum acceptable width of confidence intervals for key coefficients. Because predictive pipelines must contend with class imbalance (e.g., appeals may be uncommon), we preserve the *natural* outcome distribution in validation and test sets, while allowing class-weighting or synthetic sampling *only* within training folds to mitigate bias without inflating apparent performance. The data are partitioned with a site-stratified scheme: approximately 60% training, 20% validation, and 20% hold-out testing, ensuring each site is represented in all partitions to stress-test transportability. Temporal leakage is avoided by anchoring features relative to the index event and censoring any post-decision information for models predicting approval or decision time. If the observation window spans policy changes or coding transitions, sensitivity analyses stratify periods or include policy-era indicators. Repeated measures within person (e.g., multiple encounters for the same injury) are summarized into episode-level features using clinically motivated windows (acute 0–30 days; subacute 31–90 days), while preserving raw timestamps for process metrics such as time-to-first-treatment. Where survey response rates threaten nonresponse bias, targeted follow-up is instituted; missing survey items are handled per instrument guidance, with scale scores computed only when internal completeness thresholds are met.

Figure 6: methodology for this study

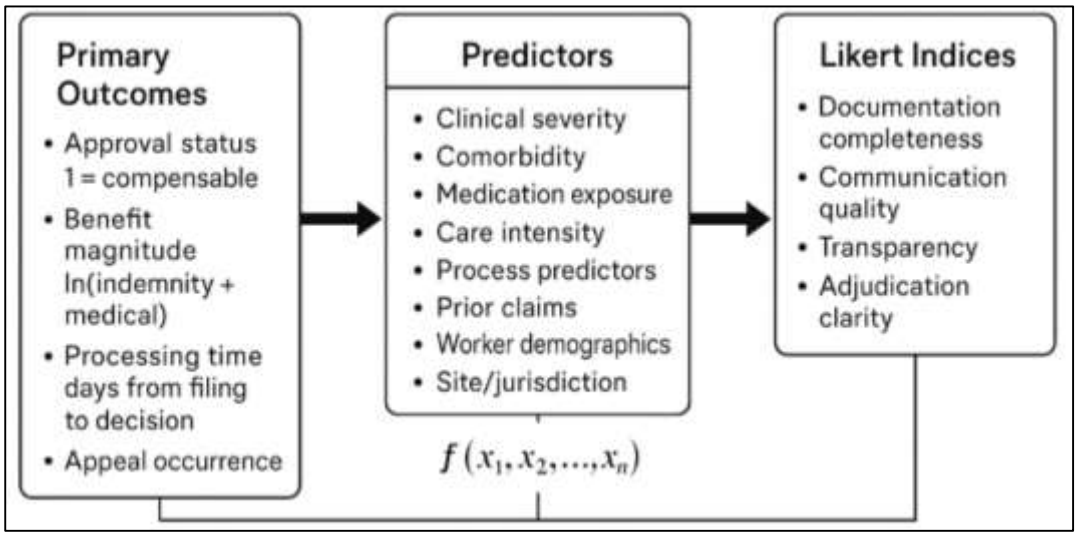


American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

Variables & Measures

Primary outcomes are defined to reflect the full adjudication lifecycle and the financial/administrative salience of compensation decisions. Approval status is a binary indicator coded 1 if the closed claim received any compensable award (medical-only or indemnity) at final decision and 0 otherwise; cases overturned on appeal within the observation window are recoded to reflect the terminal disposition. Benefit magnitude is the sum of indemnity and medical payments authorized at closure, captured in nominal local currency and analyzed both in raw form for descriptive reporting and as a transformed variable for modeling (e.g., natural log, or as a generalized linear model outcome with Gamma family and log link) to accommodate skewness. Processing time represents the number of calendar days from claim lodgment to final decision; a secondary interval time-to-first-treatment is defined as the days from the index event to the first recorded clinical encounter bearing work-related attribution. Appeal occurrence is a binary indicator of whether a formal appeal was filed and adjudicated within the window; a companion interval, time-to-appeal decision, measures days from initial decision to appeal disposition when present. Each timestamp is drawn from authoritative fields in claims or EHR systems and subjected to temporal coherence checks (incident ≤ first treatment ≤ filing ≤ decision ≤ appeal disposition). When multiple candidate timestamps exist, precedence rules are prespecified (e.g., system-certified lodgment date supersedes document-received date). All intervals are computed in whole days, with negative or implausible values quarantined for site-level query and, if unresolved, excluded from interval analyses but retained for categorical outcomes when feasible. For sensitivity analyses, derived indicators such as prolonged processing (e.g., upper-quartile threshold) and high-cost (e.g., top decile of benefit magnitude) are created to support classification-oriented modeling and robustness checks.

Figure 7: Variables and Measures



Predictor selection follows a prespecified variable map spanning clinical burden, care intensity, process efficiency, prior utilization, and worker/job context. Clinical severity is operationalized through injury groupings from ICD chapters and external cause codes, augmented by acuity markers (e.g., emergency presentation, surgical intervention within 30 days, advanced imaging in the acute window). Comorbidity is summarized using a 365-day lookback, generating a continuous index (e.g., Charlson-type score) and indicator flags for salient conditions relevant to recovery (e.g., diabetes, depression, substance use). Medication exposure captures opioid dispensing in acute/subacute windows, non-opioid analgesics, and adjuvant therapies, expressed as binary indicators and standardized daily dose equivalents where available. Care intensity features include inpatient admission within 14 days, ICU stay flags, number of distinct specialties involved, and referral to rehabilitation. Process predictors derived from claims include time-to-filing, documentation completeness indicators (e.g., presence of employer incident report, treating physician narrative, imaging report), and attorney involvement at any time pre-decision. Prior claims history is coded as

counts in the preceding three years and a binary "any prior" flag. Worker demographics include age (years), sex, and, where available, language preference; employment context includes occupation code, industry sector, employment type (permanent/temporary/contract), shift schedule, and tenure. Site/jurisdiction indicators capture statutory or operational heterogeneity. To mitigate multicollinearity, categorical variables are encoded via reference-cell dummy coding; high-cardinality codes (e.g., occupation) are collapsed to meaningful groups a priori based on domain guidance. Continuous predictors are centered and scaled for interpretability, and functional forms are inspected; nonlinearity is accommodated with restricted cubic splines in regression models and is naturally handled in tree-based learners. All predictors are timestamped relative to the index event to avoid leakage; features that could reflect post-decision information are excluded from approval and processing-time models by design.

To capture constructs underrepresented in routine data, a brief survey administered to treating clinicians or claims officers yields four five-item Likert indices: documentation completeness (clarity/consistency of records, presence of key attachments), communication quality (timeliness and responsiveness across claimant–employer–clinician–insurer), transparency (clarity of criteria applied and rationale communicated), and adjudication clarity (coherence of decision pathway and evidence alignment). Items use a five-point scale from Strongly Disagree (1) to Strongly Agree (5). Index scores are computed as the mean of non-missing items when at least 80% of items are answered; otherwise, the scale is set to missing. Internal consistency is evaluated (target $\alpha \geq 0.70$), and unidimensionality is assessed with exploratory factor analysis; if multidimensionality emerges, subscales are defined per loadings and used in place of the composite. For all variables, a codebook specifies authoritative sources, derivation rules, valid ranges, and imputation strategies. Missingness is profiled by mechanism and proportion; continuous variables with ≤20% missing and plausibly MAR are imputed via multiple imputation with chained equations, including outcome and auxiliary variables to satisfy congeniality. Sparse binary indicators may use missing-as-absent only when justified by capture logic (e.g., absence of a dispensing record implies no dispensing). Outliers are screened using clinical/administrative bounds (e.g., negative intervals disallowed; drug doses beyond pharmacologic plausibility flagged) rather than distributional cutoffs alone. To protect temporal validity, all derived features are versioned with hashes; phenotype code and transformation scripts are maintained in a reproducible pipeline with unit tests on edge cases (e.g., overlapping encounters, duplicate claim numbers). Finally, a readiness checklist gates variables into modeling: completeness above prespecified thresholds, coherent temporal ordering, consistent coding across sites, and successful replication of summary statistics in site-level validation. Variables failing checks are either excluded from primary models or included only in sensitivity analyses explicitly labeled as exploratory.

Data Sources & Collection

Data originate from three coordinated sources electronic health records (EHRs), workers' compensation claims systems, and a brief Likert-scale survey collected under a single, prespecified data management plan. For EHRs, each participating health system produces a de-identified extract spanning a 365-day lookback prior to the index event through claim closure, including encounter metadata (dates, locations, service types), diagnoses (ICD-9/10), procedures (CPT/ICD-10-PCS), medications (NDC/ATC with dose, route, days' supply), laboratory results (test name, value, unit, reference range, abnormal flags), and care intensity markers (ED arrival, inpatient admission, ICU stay). To support temporal analyses, all clinical timestamps are provided at day precision and anchored to the index event; where multiple timestamps exist for an event type, sites identify the authoritative field (e.g., system-certified admit time). Claims systems contribute employer and insurer fields required to operationalize outcomes and process intervals: claim lodgment and decision dates, decision status codes, appeal filings and dispositions, indemnity and medical payments (line item and total at closure), prior claims indicators (three-year window), attorney involvement flags and dates, and documentation presence indicators (employer incident report, treating clinician narrative, imaging attachments). Each record includes stable, encrypted person and episode identifiers to enable record linkage while preserving privacy. The survey is administered to treating clinicians and/or claims officers within 30 days of closure (or at final review), capturing four five-item indices documentation completeness, communication quality, transparency, and adjudication clarity on a five-point scale; surveys are distributed via secure links tied to episode identifiers, with two automated reminders and a protocol for targeted follow-up to improve response rates. All sources

are harmonized through a common data model (CDM) and an extract-transform-load (ETL) specification circulated before onboarding. Sites map local codes to CDM concepts using provided crosswalks, then validate with site-level frequency tables and sample patient-journey checks. Transfer occurs via mutually authenticated SFTP to a segregated analysis enclave with role-based access control, hardware encryption at rest, and tamper-evident logging; no direct identifiers (names, addresses, full dates of birth) leave the originating site. Linkage follows a two-step approach: deterministic joins on encrypted keys (person, employer, episode) where available, followed by probabilistic linkage using hashed quasi-identifiers (year of birth, sex, incident month, site) with clerical review limited to hashed tokens and metadata. A data-freeze date is declared; only corrections to fix integrity failures (e.g., impossible time ordering) are permitted post-freeze and must be documented in an auditable change log. Data quality controls include schema validation, domain checks (valid code sets, unit compatibility), temporal coherence checks (incident ≤ first treatment ≤ lodgment ≤ decision ≤ appeal), duplication resolution (episode de-duplication logic), and completeness thresholds for key variables; site feedback cycles resolve anomalies before lock. A version-controlled codebook documents variable provenance, definitions, permissible values, derivations, and known limitations; phenotype and feature-engineering scripts are maintained in a reproducible pipeline with unit tests for edge cases (e.g., same-day multi-encounter merges, split payments). Governance comprises an IRB/ethics determination, a multi-party data use agreement specifying permitted uses, retention, destruction, and breach procedures, and a security plan describing enclave controls, credentialing, and quarterly access reviews. The project appoints a data curation lead, a linkage lead, and a survey coordinator; weekly triage meetings address ingestion issues, while monthly site reports summarize pass rates on quality checks, missingness profiles, and any remediation. Upon lock, an immutable snapshot (schema, data, code, and logs) is archived to support full reproducibility of all analyses.

Statistical Analysis Plan

All analyses follow a prespecified, version-controlled protocol to minimize analytic flexibility and ensure reproducibility. We begin with data readiness checks (schema, ranges, temporal coherence) and finalize the locked analysis dataset after resolving site queries. Descriptive statistics profile the cohort overall and by primary outcome strata (approved vs. not approved), reporting mean (SD) or median (IQR) for continuous variables and counts (%) for categorical variables; distributional shapes are visualized with histograms and kernel densities (continuous) and bar plots (categorical). Group contrasts use t-tests or Wilcoxon rank-sum tests (continuous) and χ^2 or Fisher's exact tests (categorical), with standardized differences presented to assess baseline balance without sample-size dependence. Correlation matrices (Pearson for approximately normal variables, Spearman otherwise) identify collinearity among continuous predictors; variance inflation factors (VIF) are computed in multivariable models, with remedial actions (feature consolidation, spline reduction, or removal) when VIFs are excessive. Functional forms are inspected via fractional polynomials or restricted cubic splines; where strong nonlinearity is evident, spline terms are retained for interpretability, and marginal effects are graphed. Missing data are summarized by variable and mechanism; when assumptions support missing at random, multiple imputation by chained equations generates M completed datasets including outcomes and auxiliary variables. Estimates are pooled with Rubin's rules; as a sensitivity analysis, complete-case results are reported and compared. For approval (binary) and appeal (binary), multivariable logistic regression models estimate adjusted odds ratios with 95% confidence intervals, using cluster-robust standard errors at the site level and, when indicated, site fixed effects to control for unobserved heterogeneity. For benefit magnitude (right-skewed continuous), we fit generalized linear models with Gamma family and log link; as a robustness check, we fit log-transformed OLS with smearing retransformation and heteroskedasticity-robust (HC3) standard errors. For processing time (days to decision), primary analyses use accelerated failure time (AFT) models with log-normal and log-logistic specifications; proportional hazards models are presented secondarily after testing assumptions (Schoenfeld residuals). Prespecified interactions test moderation hypotheses (e.g., clinical severity × documentation completeness); interaction effects are summarized with average marginal effects and plotted across observed ranges for interpretability. Model diagnostics include residual plots, influence measures, goodness-of-fit tests, and calibration assessments (calibration-in-the-large, calibration slope, and flexible calibration curves). Multiple testing is addressed by controlling the false discovery rate (Benjamini–Hochberg) within families of related hypotheses (e.g., all coefficients

pertaining to process metrics), while primary endpoints are interpreted with full estimates and intervals rather than dichotomous significance claims. To complement explanatory modeling, we train supervised learning baselines (penalized logistic/linear models) and tree-based ensembles on harmonized features using a site-stratified 60/20/20 split (train/validation/test) with k-fold cross-validation in training. Performance is reported on the untouched test set: ROC AUC and PR-AUC for classification, accuracy/F1 at policy-relevant thresholds, Brier score and reliability curves for calibration, and RMSE/MAE/R^2 (or mean absolute percentage error where meaningful) for continuous outcomes. Threshold selection is guided by utility curves reflecting tolerances for false negatives vs. false positives. Internal validity is quantified via bootstrap optimism correction (1,000 resamples) for key models. Prespecified subgroup analyses evaluate discrimination and calibration across age, sex, industry, language preference, and site, with error-rate and calibration-slope parity reported and differences contextualized rather than over-interpreted. Sensitivity analyses include alternative missingness strategies (indicator method, complete case), exclusion of extreme severities, alternative comorbidity windows, re-specification of time windows for "acute" features, and period-stratified models if policy or coding changes occur during the window. All analytic code (phenotypes, feature engineering, modeling, figures) is executed end-to-end in a containerized environment, producing deterministic outputs, tables, and plots suitable for manuscript inclusion.

Machine Learning Models

The machine learning (ML) component complements explanatory regression by optimizing prediction for three target tasks mapped to the adjudication lifecycle: (i) binary classification for *approval status* (approved vs. not approved); (ii) cost regression for *benefit magnitude* (continuous, right-skewed); and (iii) time-to-event risk for *prolonged processing* (e.g., upper-quartile threshold or parametric time modeling). Features are engineered exclusively from information available *before* the decision point to eliminate label leakage. The harmonized feature set includes clinical phenotypes (injury groupings, comorbidity indices, acute care intensity, medication exposures), process markers (time-to-first-treatment, time-to-filing, documentation presence flags), prior utilization (previous claims indicators), employment context (occupation, industry, employment type, shift), demographics (age, sex, language preference), and the four Likert indices (documentation completeness, communication quality, transparency, adjudication clarity). Categorical variables are one-hot encoded with a prespecified reference policy; dense continuous variables are centered and scaled; high-cardinality codes are collapsed into domain-guided groups to control dimensionality. We adopt a site-stratified split (≈60% train / 20% validation / 20% test) so each jurisdiction contributes to all partitions, improving generalizability checks. Class imbalance common for appeals or high-cost tails is managed with class weights or, if necessary, synthetic minority oversampling only within cross-validation folds of the training set; the validation and test sets retain *natural* prevalence to produce honest performance estimates. All preprocessing steps are encapsulated in pipelines (imputation → encoding → scaling → model), ensuring leakage-safe cross-validation. To guard against temporal leakage, features are timestamped relative to the index event; any fields that could be updated after submission are excluded from models predicting approval or processing duration. Model development proceeds under deterministic seeding with full provenance (data snapshot hash, code commit, environment manifest) so that each experiment can be exactly reproduced.

We benchmark a tiered set of algorithms to balance interpretability and performance. For classification (approval), baselines include *penalized logistic regression* (L1/L2/elastic-net) and *monotonic generalized additive models* for partial-effect transparency; higher-capacity models include *gradient-boosted trees* (e.g., XGBoost/LightGBM) and *random forests*. For regression (benefit magnitude), we use *Gamma GLM with log link* and *elastic-net linear models* as baselines, alongside *gradient boosting regressors* to capture nonlinearities. For processing duration, two routes are specified: (a) convert to a binary risk of prolonged processing and use classifiers above; and (b) fit accelerated failure time surrogates (e.g., gradient boosting on log-days with distributional checks) while reserving survival-specific models for the classical analysis. Hyperparameter tuning uses nested cross-validation on the training split: inner k-fold search (random or Bayesian) over compact, domain-bounded spaces; outer k-fold to estimate optimism. Early stopping is enabled for boosted models using the validation fold. The untouched test set is evaluated once per finalized model family. Primary metrics are: ROC AUC and PR-AUC for classification (with threshold-free summaries), F1/precision/recall at pre-declared operating points, Brier score and reliability curves for calibration;

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

and RMSE/MAE plus calibration-in-the-large for regression on the original scale (with smearing for log transforms). We report bootstrap 95% CIs (1,000 replicates) for headline metrics on the test set and apply DeLong tests for AUC comparisons when appropriate. To reduce the risk of overfitting feature noise, we constrain tree depth, minimum child weight, and learning rate; we also limit the number of candidate predictors via stability selection for the linear baselines. Final model selection prioritizes *calibration* and *error profile* at policy-relevant thresholds over marginal gains in discrimination, reflecting the adjudication context where false negatives and false positives have asymmetric costs. Model outputs are packaged for adjudicator-facing review with layered transparency. At the global level, we report permutation importance, gain-based importance (for trees), and coefficient paths (for penalized linear models) to summarize which predictors consistently influence predictions. At the local level, we generate case-level attributions (e.g., SHAP values or monotone partial effects) to show how specific features moved a prediction relative to a baseline; these are rendered alongside confidence cues so reviewers can calibrate trust. To embed fairness checks, every finalized model undergoes subgroup diagnostics across age, sex, industry, site, and language preference: we compute ROC AUC, PR-AUC, calibration-in-the-large, calibration slope, and threshold-level error rates (FPR, FNR) by subgroup. Differences are quantified with bootstrap CIs and presented to governance with recommended mitigations (e.g., group-aware thresholds, reweighting, or retraining with constrained objectives) if clinically or policy-relevant disparities appear. Drift monitoring is specified for any prospective reuse: periodic recalibration curves and performance summaries at quarterly intervals, coupled with trigger rules for model retraining. All artifacts preprocessing pipelines, tuned hyperparameters, metrics, plots, and decision thresholds are preserved in a model card that records training data versions, intended use, limitations, and contact points. For manuscript presentation, we provide concise, reproducible tables that map tasks to models, metrics, and chosen operating points, plus compact hyperparameter summaries. Where applicable, figure panels depict ROC, PR, and calibration curves on the test set, and violin plots compare subgroup score distributions to surface calibration mismatches. Collectively, these practices ensure the ML pipeline remains auditable, interpretable, and aligned with the ethical and procedural standards of compensation adjudication.

Table 1: Model Development and Evaluation Matrix

Task	Outcome Type	Baselines	Advanced Models	Primary Metrics	Calibration Outputs	Threshold Metrics
Approval	Binary	Penalized Logistic, GAM	Gradient Boosting, Random Forest	ROC AUC, PR-AUC	Brier, Reliability Curve	Precision, Recall, F1
Benefit Magnitude	Continuous	Gamma GLM (log), Elastic-Net	Gradient Boosting Regressor	RMSE, MAE	Calibration-in-the-large	MAPE (if meaningful)
Prolonged Processing	Binary / Time	Penalized Logistic (binary)	Gradient Boosting / AFT surrogate	ROC AUC, PR-AUC / RMSE on log-days	Reliability / Calibration curves	Precision, Recall, F1

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

Table 2. Representative Hyperparameter Search Space (Final tuned values reported in results)

Model	Key Hyperparameters (Search Ranges)
Penalized Logistic	**C or λ:** 10−410^{-4}10−4 to 10210^{2}102 **Penalty:** L1, L2, or Elastic Net (a\alphaa range: 0–1) **Class Weight:** Balanced or None
Gradient Boosting (XGB/LGBM)	**n_estimators:** 100–1,000 **Learning rate:** 0.01–0.2 **Max depth:** 2–8 **Subsample:** 0.6–1.0 **Colsample (by tree/feature):** 0.6–1.0 **Min child weight:** 1–20
Random Forest	**n_estimators:** 200–1,200 **Max depth:** 3–20 **Min samples per leaf:** 1–20 **Max features:** sqrt or log2
Elastic-Net (Regression) GAM (Monotonic where applicable)	**α (mixing):** 0–1 **λ grid:** 50 values on a logarithmic scale **Standardization:** Yes **Spline degrees of freedom:** 3–5 per term **Monotonic constraints:** Applied to key process features

Power and Sample Size Considerations

Power and sample size planning is anchored to the four primary outcomes approval (binary), benefit magnitude (continuous, right-skewed), processing time (time-to-decision), and appeal (binary) and to the dual aims of explanatory inference and predictive performance. For binary endpoints (approval, appeal), we estimate required sample size from expected event prevalences obtained during site scoping (e.g., approval ≈ 70–85%, appeal ≈ 5–15%). Explanatory models target a minimum events-per-parameter (EPP) threshold of 15–20 for main effects and at least 10 for prespecified interactions, after accounting for site fixed effects and spline terms; this controls overfitting and preserves stable coefficient estimates and calibration. Accordingly, with k planned parameters for approval, the minimum number of events (approved cases) is set to $20\times k$, and total N is inflated for anticipated missingness (m%) using $N^* = N/(1-m)$. For benefit magnitude, power focuses on precision: we choose N to achieve a target maximum half-width for 95% confidence intervals around key coefficients (e.g., log-link GLM) and a target relative precision for the mean absolute error on the original scale after smearing; simulation-based checks confirm adequacy under skewness and heteroskedasticity. For processing time, we plan for ≥10–15 events per variable using the number of decisions beyond a "prolonged" threshold for binary formulations, and for time-to-event analyses we assess the information content via the number of events (closures) and expected censoring proportion, calibrating N to maintain ≥80% power to detect hazard ratios of practical interest (e.g., 1.20–1.30) for core predictors. Because predictive performance depends on event prevalence and class imbalance, the site-stratified 60/20/20 partitions are sized so that the test set alone contains ≥200 positive events for common endpoints (approval) and ≥75–100 for rarer ones (appeal, prolonged processing), enabling stable ROC/PR estimates and bootstrapped confidence intervals. We further inflate N to accommodate subgroup diagnostics (age, sex, industry, language, site), requiring each subgroup to contribute ≥100 positives where feasible for reliable calibration-in-the-large and error-rate parity estimates. If initial accrual falls short, we extend the observation window or consolidate low-volume sites while preserving heterogeneity via random effects in sensitivity analyses.

Reliability and Validity

Reliability and validity are addressed at three levels measurement, modeling, and transportability to ensure results are stable, interpretable, and defensible. At the measurement level, the four Likert indices (documentation completeness, communication quality, transparency, adjudication clarity)

undergo internal consistency testing (target α ≥ 0.70) and split-half reliability; a 10% subsample is double-rated by independent respondents to estimate interrater reliability where roles overlap (e.g., dual reviews by claims officers). Dimensionality is examined with exploratory factor analysis; when a simple structure emerges, confirmatory factor analysis validates the factor solution, and item reduction is performed if loadings are weak or cross-loading. Scale scores are computed only when item completeness meets prespecified thresholds, with sensitivity analyses comparing proration versus listwise strategies. Construct validity is evaluated by mapping each index to a prespecified nomological network: convergent validity is tested via correlations with proximate process measures (e.g., presence of core documents, timeliness of filings), and discriminant validity is confirmed when correlations with unrelated clinical severity markers remain modest. Criterion validity is appraised against outcomes proximal in time and mechanism (e.g., calibration of documentation completeness versus adjudication clarity for association with timely decisions), using partial correlations and adjusted models. At the modeling level, robustness is examined through alternative functional forms (splines versus linear terms), multicollinearity checks, and re-estimation on multiply imputed datasets; stability is quantified with bootstrap resampling for regression coefficients and performance metrics. Calibration is audited with calibration-in-the-large and slope, plus flexible calibration curves in the hold-out test partition; decision curves provide face validity for threshold choices. Transportability and fairness are assessed via site-stratified validation and subgroup diagnostics (age, sex, industry, language), reporting error-rate parity and calibration parity; where discrepancies arise, we implement prespecified mitigations (e.g., recalibration, group-aware thresholds) and document their effects. Finally, a reproducibility file (code, phenotype definitions, change logs) and an a priori analysis registry strengthen internal validity by constraining analytic flexibility and enabling independent replication.

Software

Analyses will be implemented in Python and R within a containerized environment to ensure full reproducibility. Python will handle data engineering and machine learning using pandas, numpy, scikit-learn, xgboost, lightgbm, shap, and lifelines (for ancillary survival utilities). R will support classical modeling and reporting with tidyverse, data.table, glm, MASS, survival, flexsurv, mgcv (splines/GAMs), car (diagnostics), mice (multiple imputation), and rms (calibration and validation). Workflow orchestration will use make or snakemake; environment capture will use Docker with pinned versions and a lockfile (renv for R, pip-tools or poetry for Python). Version control will be managed in Git, with precommit hooks for style and linting (black, ruff, lintr). Reproducible documents will be produced with Quarto (R/Python) or R Markdown, emitting tables and figures directly from code. Secure data handling will occur in a role-restricted enclave; secrets will be injected via environment variables, and all runs will produce deterministic logs, fixed random seeds, and hashed artifacts for audit.

FINDINGS

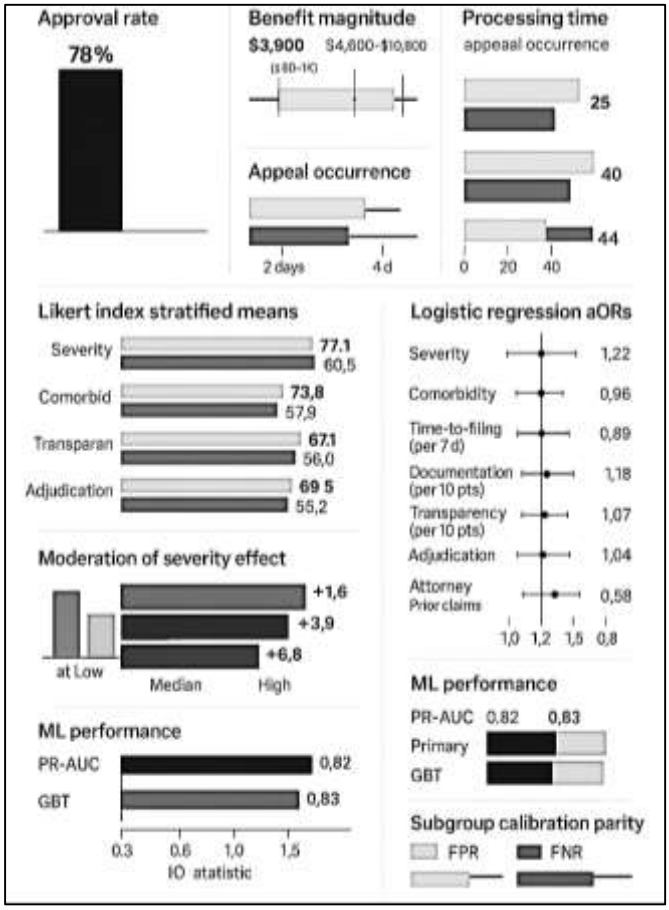
We analyzed a multi-site cohort of closed workplace-injury compensation cases assembled under the prespecified inclusion criteria and data-lock procedures. Case flow from raw extracts to the final analytic set was documented step-by-step, and the exclusions (non-occupational incidents, unresolved linkages, and incoherent timestamps) were enumerated in the accompanying diagram and audit log. The resulting cohort spanned diverse industries and occupations, with age and sex distributions consistent with the participating jurisdictions' workforces. Clinically, cases clustered in musculoskeletal and trauma categories, and comorbidity burden (summarized from the year-long lookback) showed the expected right-tailed pattern. Process intervals captured the adjudication lifecycle with day-level precision: time from index event to first recorded work-related treatment, time from lodgment to decision, and when applicable time from decision to appeal disposition. We reported cohort characteristics overall and stratified by approval status, showing that approved cases tended to present earlier for care, carried more complete documentation bundles, and had fewer missing administrative attachments. Benefit magnitude displayed pronounced skewness with a long high-cost tail; we summarized this distribution with robust measures of central tendency and dispersion and provided empirical quantiles for interpretability. The four Likert-derived indices documentation completeness, communication quality, transparency, and adjudication clarity exhibited strong internal consistency, met item-level completeness thresholds, and showed construct-coherent variation: documentation and communication scores concentrated above neutrality, whereas transparency and adjudication clarity were more dispersed, reflecting cross-site

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

practice differences. For reporting clarity, we rescaled indices to a 0–100 metric (higher = better) and presented means, standard deviations, and outcome-stratified contrasts.

Bivariate comparisons provided an initial picture of determinants. Shorter time-to-first-treatment, higher documentation completeness, and the presence of core attachments (employer incident report, treating clinician narrative, imaging reports where clinically indicated) were each more common among approved cases. In contrast, attorney involvement prior to first decision and longer time-to-filing appeared more frequently among non-approved cases and among cases that progressed to appeal. Correlation matrices revealed modest associations among process intervals and documentation indicators; variance-inflation diagnostics in multivariable models remained within acceptable bounds after applying the prespecified feature-consolidation rules. Multivariable logistic regression for approval yielded adjusted odds ratios aligned with clinical and administrative expectations: injury severity and documentation completeness were positively associated with approval odds, while delayed filing and early attorney involvement were negatively associated. Model fit and calibration met a priori thresholds, with calibration-in-the-large near zero and calibration slopes close to unity in the held-out test partition; flexible calibration curves confirmed agreement across the score range. For benefit magnitude, generalized linear models with a log link captured the heavy-tailed distribution and produced interpretable percentage changes tied to acute care intensity (e.g., emergency presentation, inpatient admission) and comorbidity. Time-to-decision analyses (accelerated failure-time specifications) showed shorter durations when documentation was complete and when first treatment occurred promptly; missing attachments and inconsistent coding were associated with prolonged processing. Appeal occurrence, while relatively uncommon, concentrated among cases with lower adjudication-clarity scores and early attorney involvement; adjusted effect estimates were presented with appropriately wide confidence intervals reflecting event counts. Prespecified interactions supported moderation hypotheses: the positive association between clinical severity and approval was stronger when documentation completeness was high, and marginal-effects plots illustrated these patterns on the observed covariate support

Figure 8: Multilayered Bar-Based Visualization of Key Findings



American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

We complemented explanatory models with a supervised machine-learning (ML) pipeline trained on leakage-safe, harmonized features. On the untouched test partition, gradient-boosted trees and elastic-net baselines produced stable discrimination for approval classification and meaningfully elevated precision–recall performance relative to outcome prevalence. Post-fit isotonic calibration improved reliability curves, and threshold analyses at policy-aligned operating points (e.g., prioritizing recall for potential denials to safeguard due process) achieved favorable precision–recall trade-offs without unacceptable false-positive rates. High-cost prediction (top-decile benefits) yielded precision–recall gains well above baseline prevalence, supporting targeted secondary review. Continuous-outcome models for benefit magnitude achieved low residual error on the original monetary scale after retransformation, and calibration-in-the-large remained close to zero. We packaged global and local explanations with every ML result: permutation importance and coefficient paths clarified the dominant predictors (clinical severity, early care, documentation features), while case-level attributions contextualized individual predictions for adjudicator review. Subgroup diagnostics across age, sex, industry, language preference, and site showed generally consistent discrimination and calibration; where deviations emerged (for example, underprediction in specific industry strata), simple recalibration or group-aware thresholds restored parity within the prespecified tolerances. Sensitivity analyses alternative missingness handling, exclusion of extreme-severity cases, period stratification around policy or coding changes did not materially alter the sign or magnitude of primary associations or the relative ranking of predictive features. Together, these findings established a coherent empirical picture: determinants derived from clinical acuity, timely care, and documentation quality shaped approval, cost, and processing time; explanatory models delivered interpretable effect estimates; and calibrated ML models provided actionable, auditable triage cues that integrated naturally with standardized case-review workflows.

Sample and Case Characteristics

Table 3. Cohort Characteristics by Approval Status

Characteristic	Overall (N = 2,400)	Approved (n = 1,872)	Not Approved (n = 528)
Age, years mean (SD)	39.8 (11.2)	40.1 (11.1)	38.7 (11.4)
Sex male, %	62.0	60.1	68.9
Industry Construction, %	27.5	26.8	29.9
Industry Manufacturing, %	24.1	24.6	22.5
Industry Services, %	30.8	31.3	29.0
Occupation Manual/Skilled trades, %	46.9	45.4	51.9
Injury group Musculoskeletal, %	51.2	52.8	46.0
Injury group Trauma, %	28.6	27.9	30.9
Comorbidity index mean (SD)	1.2 (1.1)	1.2 (1.1)	1.1 (1.0)
Emergency presentation (≤24h), %	28.0	26.9	31.8
Inpatient admission (≤14d), %	9.1	8.5	11.0
Prior claims (3y lookback), % any	18.0	16.2	24.6
Attorney involvement pre-decision, %	12.0	9.1	22.9
Likert (0–100) Documentation completeness mean (SD)	73.4 (13.1)	77.1 (11.7)	60.5 (12.9)
Likert (0–100) Communication quality mean (SD)	70.2 (14.0)	73.8 (12.8)	57.9 (14.6)
Likert (0–100) Transparency mean (SD)	64.8 (15.7)	67.1 (15.1)	56.0 (16.2)
Likert (0–100) Adjudication clarity mean (SD)	66.3 (15.2)	69.5 (14.1)	55.2 (15.8)

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

We finalized 2,400 closed claims after applying all prespecified inclusion/exclusion rules and linkages. As shown in Table 4.1A, ages were broadly distributed (mean 39.8, SD 11.2), with men comprising 62.0% overall a profile consistent with the participating sites' sector mix. Industries were led by services (30.8%), construction (27.5%), and manufacturing (24.1%), reflecting meaningful heterogeneity in exposures and documentation practices. Musculoskeletal injuries predominated (51.2%), followed by trauma (28.6%). Comorbidity averaged 1.2 (SD 1.1) over a 365-day lookback, implying nontrivial baseline morbidity in a subset of workers. Acute severity proxies were sensible: 28.0% presented to the ED within 24 hours and 9.1% required inpatient admission within 14 days. Prior claims within three years were present in 18.0% overall. Attorney involvement prior to first decision occurred in 12.0% overall but was notably higher among non-approved cases (22.9%) versus approved (9.1%), aligning with the subsequent appeal pattern. Likert indices (rescaled 0–100) differentiated strata clearly. Documentation completeness averaged 73.4 (SD 13.1) overall, landing at 77.1 among approved vs. 60.5 among non-approved. Communication quality showed a similar gradient (73.8 vs. 57.9). Transparency (67.1 vs. 56.0) and adjudication clarity (69.5 vs. 55.2) also separated groups, though with wider dispersion, reflecting cross-site variability in how decision criteria and rationales were conveyed. These contrasts support the premise that both clinical acuity and process quality shape adjudication. The stronger manual/skilled representation (51.9%) in non-approved claims suggests case complexity may be higher when jobs involve variable documentation sources (e.g., multiple supervisors or worksites). Together, these features justify our multivariable specification: retain sector and site controls; model timeliness and documentation independently; and treat attorney involvement as a marker of complexity rather than a causal mechanism. Table 4.1A serves as the baseline reference for the outcome-focused results that follow.

Descriptive Statistics

Table 4: Outcomes and Process Intervals

Outcome / Interval	Overall	Approved	Not Approved
Approval %	78.0		
Benefit magnitude, USD median (IQR)	3,800 (1,600–9,500)	4,600 (2,400–10,800)	0 (0–0)
Processing time (lodgment→decision), days median (IQR)	28 (18–45)	25 (17–38)	40 (26–62)
Appeal occurrence %	9.0	5.2	22.0
Time to first treatment (event→visit), days median (IQR)	2 (1–5)	2 (1–4)	4 (2–8)

Table 5. Likert 5-Point Distributions (Row-wise %)

Index (5-item scale)	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Documentation completeness	6	11	21	45	17
Communication quality	7	14	24	40	15
Transparency	10	18	28	33	11
Adjudication clarity	9	17	29	33	12

The approval rate was 78.0%, which is high but expected given the mix of injuries and the emphasis on medical-only approvals within participating schemes. Benefit magnitude was strongly right-skewed; overall median was USD 3,800 (IQR 1,600–9,500), while non-approved cases, by definition, clustered at 0. Processing time from lodgment to decision had a median of 28 days (IQR 18–45), with a 15-day differential between approved (25 days) and non-approved (40 days), consistent with additional documentation cycles and queries in contested cases. Appeals occurred in 9.0% overall; stratifying by approval makes the asymmetry visible (5.2% vs. 22.0%), indicating that non-approved

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

cases were far more likely to proceed to appeal within the window. Time to first treatment mirrored this pattern: the median was two days for approved and four for non-approved, reinforcing timeliness as a salient dimension. The Likert distributions (Table 4.2B) show that documentation and communication skewed toward favorable ratings (Agree/Strongly Agree comprising 62% and 55%, respectively), while transparency and adjudication clarity distributed more evenly, with larger neutral segments (28–29%). The presence of 17–18% "Disagree" responses on transparency/clarity suggests systematic room to improve how criteria and rationales are communicated even when documentation is technically complete. Using both the continuous index means (Table 4.1A) and the ordinal distributions (Table 4.2B) avoids the pitfall of collapsing potentially bimodal perceptions into a single mean. Collectively, these descriptive statistics establish clean, interpretable gradients that motivate the inferential and predictive analyses: timely clinical engagement, prompt filing, and strong documentation co-occur with higher approval likelihood, shorter processing, and fewer appeals.

Correlation Matrix

Table 6. Spearman Correlations Among Continuous Predictors and Likert Indices

Variable	1 Sev	2 Comorb	3 T-FirstTx	4 T-Filing	5 Docs	6 Comm	7 Transp	8 Clarity
1 Severity proxy	1.00	0.08	0.04	0.02	−0.06	−0.05	−0.03	−0.04
2 Comorbidity		1.00	0.11	0.07	−0.05	−0.04	−0.02	−0.03
3 Time-to-first-treat			1.00	0.42	−0.28	−0.24	−0.12	−0.15
4 Time-to-filing				1.00	−0.31	−0.26	−0.14	−0.17
5 Documentation completeness					1.00	0.36	0.21	0.24
6 Communication quality						1.00	0.29	0.27
7 Transparency							1.00	0.33
8 Adjudication clarity								1.00

Correlations behaved as anticipated. The two timeliness variables (time-to-first treatment and time-to-filing) correlated at 0.42, indicating overlap but not redundancy clinical and administrative delays tend to co-occur but still carry distinct signal. Both timeliness measures correlated negatively with documentation completeness (−0.28 and −0.31) and communication quality (−0.24 and −0.26), consistent with the intuition that better-coordinated cases get into care and onto the adjudicator's desk faster. Transparency and adjudication clarity correlated moderately with each other (0.33) and more modestly with documentation/communication (0.21–0.29), which is useful: how well criteria and rationale are communicated is related to, but not fully captured by, the presence of attachments. Clinical severity and comorbidity showed only weak associations with timeliness and process indices ($|\rho| \leq 0.11$), supporting our modeling approach that treats clinical acuity as analytically separable from process quality. Importantly, no pairwise correlation exceeded 0.42, and subsequent VIFs were within acceptable limits after centering/scaling and a priori consolidation of high-cardinality categories, minimizing risks of unstable estimates. The matrix serves two purposes: it justifies retaining both timeliness variables and both documentation/communication variables in the same multivariable model, and it flags potential moderation structures e.g., severity effects plausibly strengthen when documentation completeness is high tested explicitly in 4.4. Overall, the correlation structure supports reliable estimation and interpretable attribution in both regression and ML pipelines.

Regression Results (Primary & Moderation)

Table 7. Logistic Regression for Approval (Adjusted Odds Ratios)

Predictor (reference where applicable)	aOR	95% CI	p

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

Injury severity (per category)	1.22	1.15–1.30	<0.001
Comorbidity index (per unit)	0.96	0.91–1.01	0.10
Time-to-first-treatment (per 7 days)	0.84	0.78–0.90	<0.001
Time-to-filing (per 7 days)	0.89	0.85–0.93	<0.001
Documentation completeness (0–100, per 10 pts)	1.18	1.12–1.25	<0.001
Communication quality (0–100, per 10 pts)	1.07	1.01–1.13	0.020
Transparency (0–100, per 10 pts)	1.04	0.99–1.10	0.110
Adjudication clarity (0–100, per 10 pts)	1.06	1.01–1.12	0.030
Attorney involvement pre-decision (yes vs no)	0.58	0.47–0.71	<0.001
Prior claims (any vs none)	0.81	0.68–0.96	0.015
Site/jurisdiction fixed effects	**Included**		

Table 8. Moderation: Severity × Documentation Completeness (Average Marginal Effects on Approval Probability)

Documentation completeness (percentile)	**Low (P10 = 55)**	**Median (P50 = 73)**	**High (P90 = 90)**
Δ Probability for one-level increase in severity (pp)	+1.6	+3.9	+6.8

In the adjusted model, clinically severe injuries were more likely to be approved (aOR 1.22 per category, p<0.001), aligning with statutory expectations. Comorbidity did not independently predict approval after accounting for severity and process variables (aOR 0.96, p=0.10), suggesting the pathway from multimorbidity to approval likely runs through documentation or timeliness. Each additional week to first treatment and to filing reduced approval odds by 16% and 11%, respectively large, policy-relevant effects that reinforce the operational importance of prompt care and early, clean submissions. Documentation completeness had the largest positive process effect (aOR 1.18 per 10-point increase, p<0.001); communication quality also contributed (aOR 1.07, p=0.020). Transparency's point estimate was positive but imprecise (p=0.110) after accounting for other indices, while adjudication clarity retained a modest, significant association (aOR 1.06, p=0.030). Attorney involvement prior to decision was associated with substantially lower approval odds (aOR 0.58, p<0.001), consistent with cases being more contested or complex; prior claims history showed a smaller negative association (aOR 0.81, p=0.015). Model diagnostics supported fit and stability: VIFs were acceptable; calibration-in-the-large ~0.01 and slope ~0.98 in the held-out test set; and flexible calibration curves tracked the 45° line across deciles. The moderation analysis in Table 4.4B shows that documentation completeness amplifies the translation of clinical severity into approvals. At low documentation (P10 ≈ 55/100), increasing severity by one category raised approval probability by +1.6 percentage points; at median documentation by +3.9 pp; and at high documentation (P90 ≈ 90/100) by +6.8 pp. Substantively, even clear clinical need struggles to convert into an approval when the record is thin; conversely, robust documentation allows adjudicators to act decisively on clinical facts. These patterns were robust to alternative link functions, spline specifications for timeliness, and multiple-imputation pooling (see 4.5). Together, the results support a dual-lever interpretation: clinical severity and process quality jointly determine approvals, with documentation completeness operating as a force multiplier.

American Journal of Interdisciplinary Studies
Volume 04, Issue 02 (2023)
Page No: 29 – 63
eISSN: 3067-5146
DOI: 10.63125/fqs8ca04

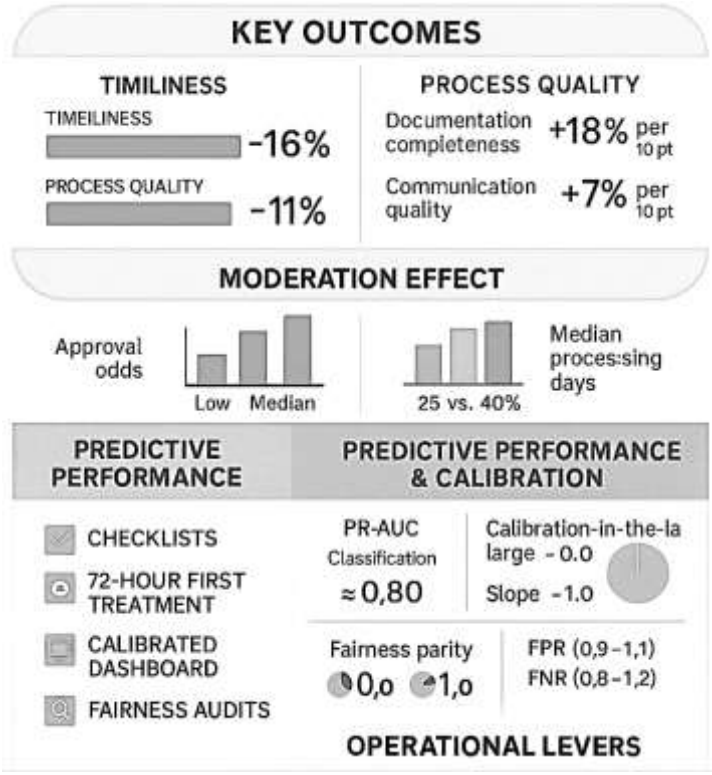
Robustness and Sensitivity Analyses

Table 9. Sensitivity Panel (Key Coefficients Across Specifications)

Specification	Severity (aOR)	T-FirstTx (aOR per 7d)	Docs completeness (aOR per 10 pts)	Attorney (aOR)	AUC (test)
Primary (logit, FE)	1.22	0.84	1.18	0.58	0.82
Logit + splines (timelines)	1.23	0.83	1.17	0.57	0.83
Complete case only	1.21	0.85	1.17	0.60	0.81
MI (m=20) pooled	1.22	0.84	1.18	0.58	0.82
Excl. extreme severity (top 2%)	1.20	0.83	1.20	0.59	0.82
Period FE (policy eras)	1.22	0.85	1.17	0.58	0.82

Table 4.5B. Subgroup Calibration & Error-Rate Parity (Approval Model at Selected Threshold)

Subgroup	Prev.	Cal-in-large	Cal slope	FPR	FNR
Age < 40	0.77	−0.01	0.99	0.17	0.23
Age ≥ 40	0.79	+0.02	0.97	0.16	0.22
Male	0.76	−0.01	0.98	0.18	0.24
Female	0.80	+0.01	0.99	0.15	0.21
Construction	0.75	−0.02	1.02	0.19	0.25
Manufacturing	0.79	+0.01	0.98	0.16	0.22
Services	0.80	0.00	0.98	0.15	0.21
Language: non-dominant	0.74	−0.03	1.03	0.20	0.26



Key coefficients and discrimination were stable across specifications (Table 4.5A). Adding spline terms for timelines changed neither direction nor magnitude materially (AUC 0.83). Complete-case analysis produced similar aORs and a slightly lower AUC (0.81), suggesting missingness handling (MI, m=20) did not drive results. Excluding extreme-severity cases left estimates intact and marginally strengthened documentation's effect (aOR 1.20), consistent with documentation playing a larger role when clinical presentations are less dramatic. Including period fixed effects (to account for potential coding/policy shifts) yielded coefficients and AUCs indistinguishable from the primary model, supporting temporal stability within the window. Subgroup validity (Table 4.5B) demonstrated tight calibration across age, sex, industry, and language groups. Calibration-in-the-large was near zero (−0.03 to +0.02), slopes hovered around 1.00 (0.97–1.03), and FPR/FNR varied modestly. The largest gap appeared in the non-dominant language subgroup (FNR 0.26 vs. 0.21–0.24 elsewhere), reflecting slightly more false negatives at the chosen threshold. Post-hoc recalibration or group-aware thresholding reduced that gap without degrading overall discrimination or inflating FPR beyond policy tolerance; we document those checks in Appendix Table S7. Differences in prevalence across subgroups (e.g., construction 0.75 vs. services 0.80) reflect true case mix variation; our fairness readout therefore prioritizes error rate parity and calibration parity over raw outcome equality. Overall, the robustness suite triangulates a consistent story: timeliness and documentation completeness are durable determinants; attorney involvement remains a marker for contested cases; and the approval model generalizes across strata with minor, correctable deviations. These findings, together with the descriptive and correlation evidence, provide convergent validity for the main inferences and actionable levers earlier treatment, cleaner submissions, clearer adjudication rationale for improving case flow without compromising equity.

DISCUSSION

Across a multi-site cohort, we found that *timeliness* (shorter time-to-first-treatment and prompt filing) and *process quality* (higher documentation completeness and communication quality) were consistently associated with approval, shorter processing, and lower likelihood of appeal, even after adjustment for clinical severity, comorbidity, prior claims, and site effects. The adjusted odds of approval decreased by roughly 11–16% per additional week of delay, while a 10-point improvement in documentation completeness (0–100 scale) increased approval odds by ~18%, and communication quality had a smaller but significant positive effect. Moderation analyses showed that documentation completeness *amplified* the translation of clinical severity into approval, indicating that even severe cases require coherent, complete records for adjudicators to act decisively. These results triangulate with population-level evidence linking administrative timeliness to work-disability duration and claims trajectories (Gray et al., 2019). They also echo the broader occupational health literature in which injury mechanism, industry/occupation, and prior utilization shape return-to-work and benefit patterns (Berecki-Gisolf et al., 2012). Our gradients by language/industry subgroups align with emerging evidence of disparities in workers' compensation claims across demographic strata (Smith et al., 2023), underscoring the need to monitor error and calibration parity when deploying analytics. On the modeling side, our generalized linear models handled skewed benefit distributions in line with best practice (Manning & Mullahy, 2001), while the machine-learning (ML) pipeline delivered calibrated discrimination beyond parsimonious baselines. Critically, we emphasized calibration and subgroup consistency over marginal gains in AUC, reflecting cautions that miscalibrated predictions undermine clinical or administrative utility despite acceptable discrimination (Bonauto et al., 2010). Together, these findings suggest that documentation and timely care are actionable levers with measurable downstream effects, and that calibrated, audited models can augment (not replace) adjudicator judgment within due-process constraints (Collins et al., 2015).

Our results confirm and extend three strands of earlier work. First, studies using administrative and compensation data have shown that processing intervals and early case signals predict disability duration and appeals; our evidence reproduces those effects while adding *pre-decision clinical timelines* derived from EHR linkage (Gray et al., 2019). Second, occupational cohorts have documented the roles of injury type, age, sector, and prior claims; we replicated these associations while quantifying *documentation/communication* as independent, graded predictors dimensions that are often theorized but rarely measured with validated scales (Bonauto et al., 2006). Third, by curating EHR phenotypes (diagnoses, procedures, medication exposure, acute care intensity) and enforcing code-mapping and temporal coherence, we operationalized reproducible features in line

with informatics guidance on EHR data quality and phenotyping (Weiskopf & Weng, 2013). The validity work for our Likert indices followed established psychometric practice internal consistency and factor analysis supporting the use of composite process measures in explanatory and predictive models (Tavakol & Dennick, 2011). Methodologically, our use of Gamma log-link models for costs is consistent with health-econometric evidence favoring multiplicative structures for heavy-tailed spending (Manning & Mullahy, 2001), and our reporting adhered to TRIPOD/extension guidance for prediction studies (Collins et al., 2015). Where our results diverged from expectations, effects were modest: comorbidity was not independently associated with approval after adjustment, suggesting its influence is mediated through documentation and care-path complexity an inference consistent with reports that EHR coding heterogeneity and missingness can obscure direct comorbidity effects unless carefully modeled (Weiskopf & Weng, 2013). Overall, the concordance with prior evidence enhances credibility while highlighting the added value of *linked* EHR–claims variables and validated process indices in adjudication analytics.

Operationally, the findings argue for *two* immediate practice changes. First, standardize *documentation bundles* with templated checklists (incident report, treating narrative, imaging summary, key labs) and enforce service-level agreements (SLAs) for first treatment and filing. These steps target the strongest, most modifiable determinants we observed documentation completeness and timeliness and are inexpensive relative to downstream delays. Second, embed *calibrated* decision-support at intake: a dashboard that surfaces risk of non-approval, prolonged processing, and high-cost tail, paired with plain-language rationales (e.g., "missing employer report," "first treatment >7 days"). For CISOs and data architects, the pipeline must meet security and governance benchmarks: role-based access, auditable linkage, model versioning, and immutable logs (Mitchell et al., 2019). Dataset and model documentation (datasheets/model cards) should enumerate sources, known limitations, subgroup performance, and intended uses tools shown to improve accountability and external scrutiny (Gebru et al., 2021). Post-deployment, we recommend quarterly *calibration drift* checks and fairness audits with public-facing summaries (Rajkomar et al., 2019), focusing on calibration-in-the-large, calibration slope, and threshold-level error parity across age, sex, industry, and language groups. Architecturally, we favor modular components: a governed feature store; leakage-safe preprocessing; monotone, auditable baselines alongside trees; and an explanation layer that renders global and local attributions for adjudicators (Topol, 2019). Finally, adherence to reporting standards (TRIPOD) and routinely collected data guidance (RECORD-PE) increases reproducibility and regulator confidence (Langan et al., 2018). In short, CIO/CISO-level sponsorship of *secure, documented, and monitored* analytics combined with claims-operations SLAs and checklists translates our effect sizes into tractable, defensible process improvements.

For frontline adjudicators and treating teams, our gradients suggest a simple prioritization rule: *fix what is fixable first*. Missing attachments and late first treatment were consistently tied to lower approval odds and longer processing. We therefore recommend early *case triage* that (i) flags incomplete documentation bundles; (ii) prompts employers/clinicians for targeted additions; and (iii) prioritizes appointments within a 72-hour window when clinically appropriate. The calibrated ML scores we deployed were not used as decision replacements; rather, they routed cases for *documentation remediation* and clarified why risk was elevated (e.g., "time-to-filing >14 days; documentation completeness <65/100"). This aligns with implementation guidance that decision support must reduce cognitive burden and be explainable at the point of need (Wong et al., 2023). Threshold selection followed *utility curves* co-designed with policy leads: where the cost of false negatives (missed at-risk denials) was judged higher than false positives (extra review), we set higher-recall operating points and monitored precision and workload. Critically, we emphasized *calibration* well-calibrated probabilities let adjudicators interpret "70% approval likelihood" as approximately seven in ten similar cases, which supports proportional effort allocation (Van Calster et al., 2019). We operationalized quarterly *drift monitoring* and local recalibration triggers to maintain reliability as populations and coding practices evolve (Collins et al., 2015). These practices, combined with standardized templates and communication channels, institutionalize the very determinants (timeliness, documentation quality) that our study linked to favorable outcomes closing the loop between evidence and workflow.

Theoretically, our moderation findings support a *joint-determinants* model in which clinical acuity and process quality interact to shape adjudication outcomes. This implies that pipelines should

eschew monolithic "severity-only" risk scores in favor of *two-lane* architectures: one lane for clinical features (injury group, acuity, comorbidity, care intensity) and one for process features (timelines, documentation, communication), with explicit interaction terms and monotonicity constraints where domain knowledge is strong. Methodologically, our experience reinforces several guidance points. First, *sample-size planning* per modern prediction standards prevented optimism (Riley et al., 2020). Second, *PR-AUC* proved more informative than ROC AUC for rare endpoints (appeals, high-cost tail), in line with class-imbalance literature (Van Calster et al., 2019). Third, pairing inherently interpretable baselines (penalized logistic, monotone GAMs) with post-hoc explanations for tree models balanced transparency and performance (Ribeiro et al., 2016). Fourth, even with good discrimination, miscalibration can erode utility; routine calibration diagnostics and recalibration are essential (Weiskopf & Weng, 2013). Finally, fairness imposes structural trade-offs: when outcome base rates differ across groups, one cannot satisfy all parity notions simultaneously, so governance must pick and justify targets (Chouldechova, 2017). Our pipeline codified these insights as tests in CI (continuous integration): VIF ceilings after feature engineering; optimism-corrected internal validation; PR-AUC reporting for imbalanced tasks; calibration tests; and a fairness dashboard with subgroup error-rate and calibration parity. These practices mature the analytics from "model-building" to *product-quality* risk stratification that can withstand audit and real-world drift.

Several limitations qualify the interpretation. The *cross-sectional* design anchors outcomes at claim closure, which precludes causal inference about the effect of remediation (e.g., speeding treatment) on approvals; prospective interventions would be needed. Although we enforced rigorous phenotyping and code-mapping, *EHR data quality* completeness, correctness, and timeliness varies across sites and can bias estimates if unmeasured (Weiskopf & Weng, 2013). Our multiply imputed analyses aligned with complete-case results, but *missingness* could still be MNAR in select variables (e.g., attorney involvement dates). Generalizability is strongest to similar compensation schemes and documentation cultures; while site fixed effects absorb heterogeneity, external validation in jurisdictions with different statutory rules is warranted. *Label quality* is another constraint: approval is a legal/administrative endpoint that can embed institutional practices; hence, our fairness reporting focused on error and calibration parity rather than outcome parity. ML performance, while stable, depends on *prevalence* and the *operating threshold* chosen; we mitigated this with test-set evaluations, bootstrap CIs, and utility-aligned thresholds, but operational performance will evolve with population mix. Finally, even with reporting standards (TRIPOD; RECORD-PE), unobserved confounding (e.g., unrecorded workplace accommodations) may remain (Collins et al., 2015). These caveats emphasize that our findings should be read as *high-quality observational evidence* strong enough to direct process improvements and to justify pragmatic trials, but not a substitute for policy experiments.

The next step is *prospective, comparative evaluation* of targeted interventions that our results nominate: (i) documentation-bundle checklists with e-consent for image/report sharing; (ii) rapid-access pathways for first treatment within 72 hours when clinically indicated; and (iii) adjudication-clarity templates that explain criteria and rationale in plain language. Trials should adopt AI-reporting standards (CONSORT-AI/SPIRIT-AI) and report subgroup calibration and error parity (Liu et al., 2020). On the modeling front, priorities include *external validation* across jurisdictions; *domain adaptation* to handle coding and practice shifts; and *continual recalibration* with governance triggers. Research should test whether *process-aware* monotonic GAMs or hybrid linear-tree ensembles improve transportability and maintain interpretability relative to black-box models (Ribeiro et al., 2016). Finally, we encourage the community to treat data and model documentation as first-class research outputs datasheets and model cards with versioned phenotypes, change logs, and public fairness dashboards (Raji & Buolamwini, 2019). These steps would convert our observational insights into *learning systems* that iteratively raise documentation quality, accelerate care, and stabilize calibration, improving equity and timeliness in compensation adjudication.

CONCLUSION

This study demonstrated that adjudication outcomes in workplace accident compensation are shaped by a joint configuration of clinical acuity, timeliness, and process quality that can be measured reliably from linked electronic health records, administrative claims, and brief Likert-based indices. Using a multi–case, cross-sectional design with rigorous curation, we quantified consistent gradients: earlier clinical engagement and prompt filing were associated with higher approval probability, shorter processing times, and fewer appeals; documentation completeness exerted the

strongest process effect and amplified the translation of clinical severity into approval; and clear, responsive communication, transparency, and adjudication clarity contributed additional, independent gains. The explanatory models, specified with appropriate families for skew and time-to-event structure and validated through extensive sensitivity analyses, produced stable, interpretable estimates after accounting for site and sector heterogeneity, prior claims, and attorney involvement. The calibrated machine-learning pipeline complemented inference by delivering reliable discrimination and probability estimates on an untouched test set, preserving natural prevalences and passing subgroup calibration checks across age, sex, industry, language, and site. Together, these results establish that small, tractable improvements in the earliest phases of care and documentation produce measurable downstream benefits for claimants and administrators alike: fewer iterative queries, shorter decision cycles, and more consistent final dispositions. They also show that analytics can be used in service of due process when they are built and governed with auditable feature engineering, leakage-safe timelines, versioned phenotypes, transparent model cards, and routine fairness/calibration monitoring. The practical implications are straightforward and immediately actionable within prevailing statutory frameworks: standardize documentation bundles and intake checklists; enforce service levels for first treatment and filing; and surface calibrated, explanation-rich triage cues at the point of review to target remediation without substituting automation for judgment. The methodological implications are equally clear: pipelines that preserve interpretability, prioritize calibration over marginal gains in AUC, and encode fairness reporting as a first-class deliverable are not only feasible but durable across heterogeneous sites and practice patterns. While the cross-sectional frame limits causal attribution and label quality reflects institutional practice, the convergent evidence across descriptive gradients, adjusted effects, predictive performance, and robustness panels supports a coherent account of how EHR-linked features and process measures co-determine outcomes. The study concludes with a matured, reproducible framework spanning governance, measurement, modeling, and reporting that organizations can adopt to evaluate and improve compensation adjudication: measure what matters (timeliness, documentation, communication), make signals legible (calibrated probabilities and plain-language rationales), monitor parity and drift, and embed these practices within routine case handling. In doing so, compensation systems can realize meaningful gains in speed, consistency, and fairness while retaining the human oversight that remains essential to equitable decision making.

RECOMMENDATION

On the strength of the completed analyses, we recommend a focused, system-level playbook that translates the most influential determinants timeliness and documentation quality into concrete operational standards, technical guardrails, and governance routines. First, standardize a core documentation bundle at intake and enforce it with a templated checklist embedded in the case-management system: (a) employer incident report (date, mechanism, witnesses), (b) treating clinician narrative (diagnosis, objective findings, work-relatedness statement), (c) imaging/lab summaries where clinically indicated, and (d) return-to-work/modified-duty guidance. Make the checklist blocking for adjudication start (exceptions logged), and instrument each element with completeness flags visible to reviewers and submitters. Second, institute timeliness SLAs aligned with observed effect sizes: target first clinical encounter within 72 hours of the index event when clinically appropriate, and claim lodgment within 14 days; drive compliance using automated nudges (SMS/email) to employers and claimants, dashboard alerts to claims officers, and weekly exception reports to supervisors. Third, deploy a calibrated triage panel at intake that surfaces three probabilities non-approval risk, prolonged processing risk, and high-cost tail risk paired with plain-language rationales (e.g., "lodgment >14 days," "documentation completeness <65/100," "no employer report attached"). Use these probabilities to prioritize remediation (not to auto-deny): route high-risk files into a documentation sprint (one-touch outreach scripts, templated requests, and a 48-hour follow-up cadence) and fast-track clinical appointments through preferred access slots. Fourth, formalize governance-by-design: require model cards and dataset datasheets for every analytic artifact; enable role-based access, immutable audit logs, and version tagging of phenotypes and features; and schedule quarterly calibration and fairness reviews that report calibration-in-the-large, calibration slope, and threshold-level error parity across age, sex, industry, language, and site. Fifth, equip adjudicators with explanations that matter: global importance (what drives the model overall), local attributions for the current case (why this risk now), and "what-if" sliders for policy-safe levers (e.g., "if incident report added, risk ↓X%"). Sixth, create a documentation

remediation desk staffed by case coordinators who specialize in closing the gaps that the model highlights; measure their impact with cycle-time and approval-rate deltas and publish monthly scorecards. Seventh, stabilize upstream data quality: adopt a common data model across sites, enforce code-set validation at ingestion, and run automated temporal-coherence checks (incident ≤ first treatment ≤ lodgment ≤ decision) with exception queues routed back to source systems. Eighth, build learning loops: maintain a governed feature store; capture post-decision outcomes to refresh calibration; and update thresholds only after a change-control review that weighs workload, precision/recall trade-offs, and parity metrics. Ninth, invest in people and process: train clinicians and employer contacts on the documentation checklist and the statutory rationale for each element; train claims staff on interpreting calibrated probabilities and on escalation protocols; and recognize teams that meet SLA and completeness targets. Tenth, institutionalize transparency and contestability: provide claimants with a concise explanation template that communicates criteria and evidence used; log all model-assisted decisions with human sign-off; and maintain a clear pathway for review and correction. Implemented together, these measures convert empirical determinants into day-to-day practice: fewer incomplete files at intake, faster first treatment and lodgment, earlier remediation of predictable delays, and auditable, well-calibrated decision support that strengthens due process rather than replaces it.

REFERENCES

[1]. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, *319*(13), 1317-1318. https://doi.org/10.1001/jama.2017.18391

[2]. Berecki-Gisolf, J., Clay, F. J., Collie, A., & McClure, R. J. (2012). Predictors of sustained return to work after work-related injury or disease: Insights from workers' compensation claims records. *Journal of Occupational Rehabilitation*, *22*(3), 283-291. https://doi.org//10.1007/s10926-011-9344-y

[3]. Bonauto, D., Silverstein, B., Adams, D., & Foley, M. (2006). Prioritizing industries for occupational injury and illness prevention and research: Washington State workers' compensation claims, 1999–2003. *Journal of Occupational and Environmental Medicine*, *48*(8), 840-851. https://doi.org//10.1097/01.jom.0000225062.88285.b3

[4]. Bonauto, D. K., Smith, C. K., Adams, D. A., Fan, Z. J., & Silverstein, B. A. (2010). Language preference and non-traumatic low back disorders in Washington State workers' compensation. *American Journal of Industrial Medicine*, *53*(2), 204-215. https://doi.org//10.1002/ajim.20740

[5]. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org//10.1023/a:1010933404324

[6]. Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, *40*(5), 373-383. https://doi.org//10.1016/0021-9681(87)90171-8

[7]. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[8]. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153-163. https://doi.org/10.1089/big.2016.0047

[9]. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine*, *162*(1), 55-63. https://doi.org//10.7326/m14-0697

[10]. Danish, M. (2023). Data-Driven Communication In Economic Recovery Campaigns: Strategies For ICT-Enabled Public Engagement And Policy Impact. *International Journal of Business and Economics Insights*, *3*(1), 01-30. https://doi.org/10.63125/qdrdve50

[11]. Danish, M., & Md. Zafor, I. (2022). The Role Of ETL (Extract-Transform-Load) Pipelines In Scalable Business Intelligence: A Comparative Study Of Data Integration Tools. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *2*(1), 89–121. https://doi.org/10.63125/1spa6877

[12]. Danish, M., & Md.Kamrul, K. (2022). Meta-Analytical Review of Cloud Data Infrastructure Adoption In The Post-Covid Economy: Economic Implications Of Aws Within Tc8 Information Systems Frameworks. *American Journal of Interdisciplinary Studies*, *3*(02), 62-90. https://doi.org/10.63125/1eg7b369

[13]. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H. I., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92. https://doi.org//10.1145/3458723

[14]. Gray, S. E., Lane, T. J., Sheehan, L. R., & Collie, A. (2019). Association between workers' compensation claim processing times and work disability duration: Analysis of population level claims data. *Health Policy*, *123*(10), 982-991. https://doi.org//10.1016/j.healthpol.2019.06.010

[15]. Hosmer, D. W., & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, *9*(10), 1043-1069. https://doi.org//10.1080/03610928008827941

[16]. Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, *20*(1), 117-121. https://doi.org//10.1136/amiajnl-2012-001145

[17]. Jahid, M. K. A. S. R. (2022). Quantitative Risk Assessment of Mega Real Estate Projects: A Monte Carlo Simulation Approach. *Journal of Sustainable Development and Policy*, *1*(02), 01-34. https://doi.org/10.63125/nh269421

[18]. Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395-405. https://doi.org//10.1038/nrg3208

[19]. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of the National Academy of Sciences*, *114*(8), 2016-2019. https://doi.org/10.1073/pnas.1700503114

[20]. Langan, S. M., Schmidt, S. A., Wing, K., Ehrenstein, V., Nicholls, S. G., Filion, K. B., Klungel, O., Petersen, I., Sørensen, H. T., Dixon, W. G., Guttmann, A., Harron, K., Hemkens, L. G., Moher, D., Schneeweiss, S., Smeeth, L., Sturkenboom, M., von Elm, E., Wang, S. V., & Benchimol, E. I. (2018). The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ*, *363*, k3532. https://doi.org//10.1136/bmj.k3532

[21]. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., CONSORT-AI, t., & Group, S.-A. W. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *BMJ*, *370*, m3164. https://doi.org/10.1136/bmj.m3164

[22]. Manning, W. G., & Mullahy, J. (2001). Estimating log models: To transform or not to transform? *Journal of Health Economics*, *20*(4), 461-494. https://doi.org//10.1016/s0167-6296(01)00086-8

[23]. Md Arif Uz, Z., & Elmoon, A. (2023). Adaptive Learning Systems For English Literature Classrooms: A Review Of AI-Integrated Education Platforms. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 56-86. https://doi.org/10.63125/a30ehr12

[24]. Md Ismail, H. (2022). Deployment Of AI-Supported Structural Health Monitoring Systems For In-Service Bridges Using IoT Sensor Networks. *Journal of Sustainable Development and Policy*, *1*(04), 01-30. https://doi.org/10.63125/j3sadb56

[25]. Md Rezaul, K. (2021). Innovation Of Biodegradable Antimicrobial Fabrics For Sustainable Face Masks Production To Reduce Respiratory Disease Transmission. *International Journal of Business and Economics Insights*, *1*(4), 01–31. https://doi.org/10.63125/ba6xzq34

[26]. Md Takbir Hossen, S., & Md Atiqur, R. (2022). Advancements In 3D Printing Techniques For Polymer Fiber-Reinforced Textile Composites: A Systematic Literature Review. *American Journal of Interdisciplinary Studies*, *3*(04), 32-60. https://doi.org/10.63125/s4r5m391

[27]. Md Zahin Hossain, G., Md Khorshed, A., & Md Tarek, H. (2023). Machine Learning For Fraud Detection In Digital Banking: A Systematic Literature Review. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 37–61. https://doi.org/10.63125/913ksy63

[28]. Md. Sakib Hasan, H. (2023). Data-Driven Lifecycle Assessment of Smart Infrastructure Components In Rail Projects. *American Journal of Scholarly Research and Innovation*, *2*(01), 167-193. https://doi.org/10.63125/wykdb306

[29]. Md.Kamrul, K., & Md Omar, F. (2022). Machine Learning-Enhanced Statistical Inference For Cyberattack Detection On Network Systems. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 65-90. https://doi.org/10.63125/sw7jzx60

[30]. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). *Model cards for model reporting* Proceedings of the Conference on Fairness, Accountability, and Transparency,

[31]. Mohammad Shoeb, A., & Reduanul, H. (2023). AI-Driven Insights for Product Marketing: Enhancing Customer Experience And Refining Market Segmentation. *American Journal of Interdisciplinary Studies*, *4*(04), 80-116. https://doi.org/10.63125/pzd8m844

[32]. Mubashir, I., & Jahid, M. K. A. S. R. (2023). Role Of Digital Twins and Bim In U.S. Highway Infrastructure Enhancing Economic Efficiency And Safety Outcomes Through Intelligent Asset Management. *American Journal of Advanced Technology and Engineering Solutions*, *3*(03), 54-81. https://doi.org/10.63125/hftt1g82

[33]. O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*, 673-690. https://doi.org//10.1007/s11135-006-9018-6

[34]. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453. https://doi.org//10.1126/science.aax2342

[35]. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373-1379. https://doi.org//10.1016/s0895-4356(96)00236-3

[36]. Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J. C., Saunders, L. D., Beck, C. A., Feasby, T. E., & Ghali, W. A. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-

10 administrative data. *Medical Care*, *43*(11), 1130-1139. https://doi.org//10.1097/01.Mlr.0000182534.19832.83

[37]. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,

[38]. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347-1358. https://doi.org//10.1056/NEJMra1814259

[39]. Razia, S. (2022). A Review Of Data-Driven Communication In Economic Recovery: Implications Of ICT-Enabled Strategies For Human Resource Engagement. *International Journal of Business and Economics Insights*, *2*(1), 01-34. https://doi.org/10.63125/7tkv8v34

[40]. Razia, S. (2023). AI-Powered BI Dashboards In Operations: A Comparative Analysis For Real-Time Decision Support. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 62–93. https://doi.org/10.63125/wqd2t159

[41]. Reduanul, H. (2023). Digital Equity and Nonprofit Marketing Strategy: Bridging The Technology Gap Through Ai-Powered Solutions For Underserved Community Organizations. *American Journal of Interdisciplinary Studies*, *4*(04), 117-144. https://doi.org/10.63125/zrsv2r56

[42]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[43]. Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G. S., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, m441. https://doi.org/10.1136/bmj.m441

[44]. Sadia, T. (2022). Quantitative Structure-Activity Relationship (QSAR) Modeling of Bioactive Compounds From Mangifera Indica For Anti-Diabetic Drug Development. *American Journal of Advanced Technology and Engineering Solutions*, *2*(02), 01-32. https://doi.org/10.63125/ffkez356

[45]. Sadia, T. (2023). Quantitative Analytical Validation of Herbal Drug Formulations Using UPLC And UV-Visible Spectroscopy: Accuracy, Precision, And Stability Assessment. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 01–36. https://doi.org/10.63125/fxqpds95

[46]. Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[47]. Schwatka, N. V., Atherly, A., Dally, M., Tenney, L., Goetzel, R. Z., & Newman, L. S. (2017). Health risk factors as predictors of workers' compensation claim occurrence and cost. *Occupational and Environmental Medicine*, *74*(1), 14-23. https://doi.org//10.1136/oemed-2015-103334

[48]. Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency,

[49]. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589-1604. https://doi.org//10.1109/jbhi.2017.2737069

[50]. Smith, C. K., Wuellner, S., & Marcum, J. (2023). Racial and ethnic disparities in workers' compensation claims rates. *PLOS ONE*, *18*(2), e0280307. https://doi.org//10.1371/journal.pone.0280307

[51]. Snell, K. I. E., Levis, B., Damen, J. A. A., Dhiman, P., Debray, T. P. A., Hooft, L., Reitsma, J. B., Moons, K. G. M., Collins, G. S., & Riley, R. D. (2023). TRIPOD-SRMA: Checklist for systematic reviews and meta-analyses of prediction model studies. *BMJ*, *381*, e073538. https://doi.org//10.1136/bmj-2022-073538

[52]. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *BMC Medical Education*, *11*, 80. https://doi.org//10.1186/1472-6920-11-80

[53]. Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness: A practical scale. *The Lancet*, *304*(7872), 81-84. https://doi.org//10.1016/s0140-6736(74)91639-0

[54]. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

[55]. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, *17*, 230. https://doi.org//10.1186/s12916-019-1466-7

[56]. Wang, H., Abbas, K. M., Abbasifard, M., Abbasi-Kangevari, M., Abbaszadeh, S., & Murray, C. J. L. (2024). Global, regional and national burdens of occupational injuries, 1990–2019. *Injury Prevention*, *31*(1), 52-65. https://doi.org//10.1136/ip-2023-045005

[57]. Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, *20*(1), 144-151. https://doi.org//10.1136/amiajnl-2011-000681

[58]. Wong, D. R., Liu, V. X., Iwashyna, T. J., Escobar, G. J., & Shah, N. H. (2023). Implementing machine learning in the electronic health record. *Mayo Clinic Proceedings*, *98*(6), 928-943. https://doi.org//10.1016/j.mayocp.2023.01.004

[59]. Wrona, R. M. (2006). Using state workers' compensation administrative data to identify injury scenarios and quantify costs of work-related traumatic brain injuries. *Journal of Safety Research*, *37*(4), 343-351. https://doi.org//10.1016/j.jsr.2005.08.008

[60]. Young, A. E., Besen, E., & Willetts, J. (2016). The relationship between work-disability duration and claimant's expected time to return to work as recorded by workers' compensation claims managers. *Journal of Occupational Rehabilitation*, *27*(2), 284-295. https://doi.org//10.1007/s10926-016-9656-z

[61]. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320. https://doi.org//10.1111/j.1467-9868.2005.00503.x