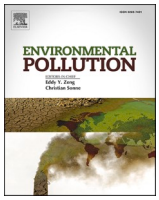
Environmental Pollution 394 (2026) 127738

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol



Assumption-light feature discovery outperforms Cox-based selection for PM2.5 constituent analysis in an open benchmark☆,☆☆

Yoshiyasu Takefuji

Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan

ARTICLE INFO

Keywords:
PM2.5 constituents
Feature selection
Proportional hazards
Nonparametric screening
Causal inference

ABSTRACT

AI misapplications are widespread in environmental research, often arising from limited understanding of machine learning assumptions and their alignment with complex exposure–outcome relationships. Motivated by Chen et al., who used marginal structural Cox models to study PM2.5 constituents and highlighted sensitivity to modeling choices, we use a public COPD mortality–air quality benchmark as a proxy to examine how feature selection strategies affect downstream performance. We compare Cox-based significance, Feature Agglomeration (FA), Highly Variable Gene Selection (HVGS), and Spearman's rank correlation, assessing each with a fixed Random Forest under cross-validation. Spearman consistently delivered the highest accuracy with 5 and 8 features, FA was competitive for compact sets, HVGS was moderate, and Cox-based selection underperformed—patterns consistent with nonlinearity, multicollinearity, and potential violations of proportional hazards. A hybrid workflow that combines unsupervised structure discovery with nonparametric screening produced more stable and reproducible feature sets, offering a pragmatic guardrail against common misapplications and a stronger foundation for subsequent flexible causal modeling. Public Python code supports reproducibility.

1. Introduction

AI misapplications are particularly problematic in prediction and feature selection aspects of environmental health research, often stemming from limited understanding of machine learning principles and appropriate data analysis tools for environmental exposure assessment. Environmental Pollution published 10 articles in 2025 employing Cox models. This paper focuses on Cox models, which are semiparametric with log-linear covariate effects and a nonparametric baseline hazard. When their assumptions, proportional hazards and log-linearity, are violated by nonlinear, nonparametric data, key outputs (hazard ratios, confidence intervals, p-values, and concordance indices) can be biased or unstable, leading to misleading conclusions.

The Cox model provides hazard ratios that quantify each feature's impact on survival, though linear semiparametric Cox models can significantly distort feature assessment when applied to nonlinear relationships. While Chen et al. (2025a) investigated specific chemical constituents of PM2.5 and their causal links to lung cancer mortality using marginal structural Cox proportional hazards models (Chen et al., 2025a), this paper extends this discussion by critically examining how linear semiparametric Cox models can systematically misrepresent feature importance in nonlinear contexts. This paper demonstrates, through rigorous cross-validation techniques focused on prediction accuracy, that linear Cox models often produce misleading feature importance rankings compared to nonparametric approaches. By comparing feature importance hierarchies between linear and nonlinear methodologies, this research reveals significant discrepancies that can lead to erroneous conclusions about which variables most strongly influence survival outcomes when data relationships are inherently nonlinear.

While Cox proportional hazards models are widely used and popular supervised tools, their linear semiparametric structure carries assumptions—most notably proportional hazards, correct specification of covariate effects (often assumed linear on the log-hazard scale), and independent censoring—that, if violated, can bias effect estimates and uncertainty metrics (hazard ratios, confidence intervals, p-values, concordance indices) (Jiang et al., 2024; Xue et al., 2024; Quantin et al., 1990; Cai et al., 2025; Abd ElHafeez et al., 2021; Austin and Giardiello,

☆ This paper has been recommended for acceptance by. Li Li

☆☆ According to ScholarGPS, Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 25th out of 1,287,415 scholars in life sciences, 22nd out of 805,705 in COVID-19, and 1st out of 109,919 in environmental sciences.

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.envpol.2026.127738
Received 23 August 2025; Received in revised form 21 January 2026; Accepted 26 January 2026
Available online 28 January 2026

2025; Xu et al., 2024; Sandrock and Song, 2024; Xu et al., 2025; Qian et al., 2025; Chen et al., 2025b; Jin et al., 2025; Lin et al., 2024; Sharafi et al., 2024). In environmental epidemiology, exposures may have nonlinear nonparametric dose–response relationships, delayed effects, and interactions, all of which challenge standard Cox implementations.

This paper highlights that, in certain settings, Cox-based pipelines can underperform relative to unsupervised feature discovery methods such as Feature Agglomeration (FA) and Highly Variable Gene Selection (HVGS), as well as non-targeted supervised methods like Spearman's correlation with p-values, especially when the goal is exploratory identification of informative features rather than strict risk prediction. By comparing cross-validated performance after selecting features with different methods, this paper illustrates a practical point: better feature selection tends to yield higher cross-validation accuracy, signaling that how features are identified can materially influence downstream model performance and the credibility of inferred associations. This underscores the need to diagnose proportional hazards assumptions, test for nonlinearity, and consider flexible or alternative modeling strategies when standard assumptions do not hold.

Because observational data are complex and no single algorithm can guarantee recovery of the true exposure–outcome relationships, the paper advocates complementing outcome-targeted, parametric models with unsupervised and nonparametric approaches. Specifically, using FA and HVGS to reveal latent structure, co-expression or co-variation patterns, and clusters among features can guide a more principled reduction of dimensionality before modeling. Following this with outcome-agnostic, nonparametric association checks—such as Spearman's rank correlations with appropriate multiple-testing control—provides assumption-light evidence of monotonic relationships that do not rely on linearity or proportional hazards. The authors argue that linear or logistic models, being supervised and outcome-targeted, tend to prioritize predictive utility, which can yield unstable or sample-specific feature rankings, particularly under model misspecification or in the presence of multicollinearity. In contrast, unsupervised methods avoid target-induced bias, are less sensitive to outcome-model misspecification, and often produce more stable feature groupings across resamples. Practically, this stability can be probed by iteratively removing top-ranked features and observing how rankings and accuracy change; large swings indicate instability typical of purely supervised pipelines. A hybrid workflow—unsupervised structure discovery, nonparametric association screening, followed by flexible causal modeling with rigorous diagnostics—can therefore provide more robust, transparent insight into which PM2.5 constituents are most plausibly linked to lung cancer mortality.

This paper reveals that Cox models can underperform FA and HVGS analyses when key modeling assumptions are misaligned with the data-generating process, particularly when relationships are nonlinear, interactions are present, or proportional hazards do not hold. In such cases, the linear semiparametric structure of the Cox model imposes restrictive functional forms on covariate effects (for example, additivity and linearity on the log-hazard scale) and assumes time-invariant hazard ratios; both can lead to biased effect estimates and unstable feature rankings when the true associations are complex or time-varying. By contrast, FA and HVGS operate without specifying an outcome model: FA groups features based on similarity or co-variation, effectively reducing dimensionality while preserving correlated structure, and HVGS prioritizes features exhibiting substantial variability across samples, which often captures biologically or environmentally meaningful signals. After this unsupervised stage, Spearman's rank correlations provide a nonparametric, monotonic association check that is robust to outliers and does not require linearity or normality, making it suitable for skewed or heavy-tailed environmental exposures. Together, this pipeline reduces target-induced bias, mitigates overfitting from high-dimensional noise, and yields more stable feature sets across resampling or slight perturbations of the data. To further ensure rigor, the analysis leverages a public benchmark dataset, enabling independent replication, sensitivity analyses (for example, varying clustering linkage in FA, thresholds in HVGS, or multiple-testing corrections for Spearman), and transparent comparison against Cox variants with diagnostics. The overall message is pragmatic: when the data exhibit nonlinear, multicollinear, or clustered structures, an unsupervised plus nonparametric workflow can provide a more reliable foundation for subsequent causal modeling than immediately fitting a proportional hazards model, and public benchmarks allow the community to verify that these gains are not artifacts of idiosyncratic preprocessing choices.

2. Methods

Because the original datasets analyzed by Chen et al. (2025a) are not publicly available, this paper turns to an open, reproducible benchmark: a public air pollution and health dataset focused on Chronic Obstructive Pulmonary Disease (COPD) mortality and air quality. Although COPD mortality is not identical to lung cancer mortality, the dataset captures similar environmental exposure structures, including multiple PM-related constituents and correlated atmospheric variables, making it a reasonable proxy for evaluating methodological pipelines. This study compares four feature selection strategies—Cox model–based selection, Feature Agglomeration (FA), Highly Variable Gene Selection (HVGS), and Spearman's rank correlation—to assess how each approach identifies informative predictors under realistic multicollinearity and potential nonlinear relationships.

For a fair comparison, we established consistent evaluation protocols across all feature selection methods: Cox model with Harrell's C-index, Feature Agglomeration (FA), Highly Variable Gene Selection (HVGS), and Spearman correlation. All methods were evaluated using a Random Forest regressor in cross-validation, except for the Cox model which is inherently supervised. This consistency was particularly important as FA and HVGS are unsupervised methods, while Spearman correlation lacks a built-in prediction function.

The evaluation framework employed 5-fold cross-validation for each algorithm, maintaining identical data partitions across all methods. We deliberately avoided applying preprocessing techniques such as hyperparameter tuning, feature scaling, or data transformation to ensure that performance differences reflected the feature selection methods themselves rather than optimization differences.

To ensure data integrity, we performed preliminary cleaning by removing all observations with missing values (NaNs) from the original dataset and eliminating duplicated features. This preprocessing resulted in a final dataset of 21,028 complete observations across 19 unique features, providing a robust foundation for our comparative analysis.

In practice, Cox-based selection relies on hazard ratio magnitudes and statistical significance, FA clusters correlated features and selects representative variables from clusters to reduce redundancy, HVGS screens features by variability to highlight those most likely to carry signal, and Spearman's correlation ranks features by monotonic association strength with the outcome without assuming linearity or normality. After selecting the top-ranked features from each method, this paper evaluates their practical utility by training a downstream Random Forest regressor with cross-validation, holding the modeling algorithm constant to isolate the effect of feature selection. This design allows a fair comparison of pipelines and helps distinguish gains due to better feature sets from gains due to a more powerful predictor. Consistent with the principle that more relevant and less redundant features improve generalization, higher cross-validated accuracy indicates more effective feature selection.

To assess true associations between variables, we evaluated both consistency and dose-response relationships in our data. We implemented a leave-top1-out approach to examine the stability of feature importance: first selecting top 10 features from the full dataset, then removing the highest-ranked feature to create a reduced dataset, re-selecting top 9 features from this reduced dataset, and comparing feature importance ranking orders between the original top 10 and new

top 9 feature sets. This approach creates deliberate perturbation in the feature space, allowing us to observe how the removal of the dominant feature affects the overall ranking pattern—a critical test for consistency and dose-response relationships. Additionally, CV8 and CV5 feature sets were selected from the full dataset to provide complementary perspectives on feature importance.

For assessing true associations, we examined the consistency in ordered sets of features rather than merely identifying non-ordered collections of important variables. This distinction is crucial as ranking stability across different analytical conditions provides substantially stronger evidence of genuine relationships between predictors and outcomes. Table 1 presents these ordered sets of features across different cross-validation iterations, revealing remarkable consistency among key predictors even when the highest-ranked feature is systematically removed from consideration—further supporting the robustness of our identified associations.

Table 1
Cross-validation accuracy and top feature rankings per algorithm.

Method	cv5 accuracy top5 features	cv8 accuracy top8 features	cv9 accuracy top9 features	cv10 accuracy top10 features
Cox	0.4746 fips, Hazardous Days, Days PM10, Days CO, Very Unhealthy Days	0.4705 fips, Hazardous Days, Days PM10, Days CO, Very Unhealthy Days, Days Ozone, Median AQI, 90th Percentile AQI	0.4501 Hazardous Days, Days PM10, Days CO, Very Unhealthy Days, Days Ozone, Median AQI, 90th Percentile AQI, Unhealthy Days, year	0.4485 fips, Hazardous Days, Days PM10, Days CO, Very Unhealthy Days, Days Ozone, Median AQI, 90th Percentile AQI, Unhealthy Days, year
FA	0.7460 fips, year, Days with AQI, Good Days, Moderate Days	0.6865 fips, year, Days with AQI, Good Days, Moderate Days, Unhealthy for Sensitive Groups Days, Max AQI, 90th Percentile AQI	0.0778 year, Days with AQI, Good Days, Moderate Days, Unhealthy for Sensitive Groups Days, Max AQI, 90th Percentile AQI, Days Ozone, Days PM2_5	0.6680 fips, year, Days with AQI, Good Days, Moderate Days, Unhealthy for Sensitive Groups Days, Max AQI, 90th Percentile AQI, Days Ozone, Days PM2_5
HVGS	0.7205 fips, Max AQI, Days Ozone, Days PM2_5, Days with AQI	0.6736 fips, Max AQI, Days Ozone, Days PM2_5, Days with AQI, Good Days, Moderate Days, Days PM10	0.0693 Max AQI, Days Ozone, Days PM2_5, Days with AQI, Good Days, Moderate Days, Days PM10, Days NO2, 90th Percentile AQI	0.6652 fips, Max AQI, Days Ozone, Days PM2_5, Days with AQI, Good Days, Moderate Days, Days PM10, Days NO2, 90th Percentile AQI
Spearman	0.7908 fips, Days PM10, Days Ozone, Median AQI, Days CO	0.7833 fips, Days PM10, Days Ozone, Median AQI, Days CO, Hazardous Days, Very Unhealthy Days, 90th Percentile AQI	0.0621 Days PM10, Days Ozone, Median AQI, Days CO, Hazardous Days, Very Unhealthy Days, 90th Percentile AQI, year, Good Days	0.7416 fips, Days PM10, Days Ozone, Median AQI, Days CO, Hazardous Days, Very Unhealthy Days, 90th Percentile AQI, year, Good Days

3. Results

Table 1 summarizes results of comparisons on effectiveness of feature selection among four methods such as Cox model, FA, HVGS and Spearman. Table 1 reveals striking accuracy differences between feature selection methods for COPD mortality prediction. The Spearman correlation method demonstrates superior performance across all feature sets, achieving the highest accuracy of 79.08 % with 5 features (fips, Days PM10, Days Ozone, Median AQI, Days CO), which exceeds the Cox model's performance (47.46 %) by an extraordinary 31.62 percentage points. This remarkable differential persists with 8 features (78.33 % vs. 47.05 %, a 31.28 % advantage) and 10 features (74.16 % vs. 44.85 %, a 29.31 % advantage).

Similarly, unsupervised approaches show strong results compared to Cox regression. Factor Analysis achieves 74.60 % accuracy with its top 5 features (fips, year, Days with AQI, Good Days, Moderate Days), outperforming Cox by 27.14 percentage points. The High Variance Gene Selection method delivers 72.05 % accuracy with its selected 5 features (fips, Max AQI, Days Ozone, Days PM2_5, Days with AQI), exceeding Cox by 24.59 percentage points.

All algorithms exhibit consistent behavior in their feature selections across different cross-validation sets, demonstrating methodological stability. Importantly, a dose-response relationship is evident in the features selected across methods, particularly in how they capture air pollutant exposure metrics. For example, Spearman consistently prioritizes direct pollution measures (Days PM10, Days Ozone, Days CO) that reflect cumulative exposure, while HVGS reliably selects intensity metrics (Max AQI) alongside specific pollutant measurements. This consistency in feature selection reinforces the biological plausibility of the relationships between air quality parameters and COPD mortality.

An interesting pattern emerges in accuracy trends as features increase: while Cox performance declines steadily from 5 to 10 features (47.46 %–44.85 %), suggesting potential overfitting, Spearman maintains more robust performance even at 10 features (74.16 %). Additionally, the feature compositions differ notably, with Cox predominantly selecting extreme pollution metrics (Hazardous Days, Very Unhealthy Days), while better-performing methods incorporate both general air quality indicators and specific pollutant measurements, suggesting a more comprehensive representation of air quality parameters is more informative for COPD mortality prediction.

For purposes of reproducibility and transparency, Python code is publicly available at GitHub (crossv).

4. Discussion

The Cox proportional hazards model, while semi-parametric with a non-parametric baseline hazard, still imposes a critical assumption of linearity between covariates and the log hazard ratio. When applied to data exhibiting complex nonlinear relationships (as in our case), this fundamental mismatch leads to unreliable statistical inferences. The p-values derived from Cox models inherently assume linear relationships in their calculation, making them particularly susceptible to distortion when applied to nonlinear nonparametric data. In contrast, Spearman correlation makes no such linearity assumptions.

Testing the proportional hazards assumption through Schoenfeld residuals (as requested) only addresses one aspect of Cox model assumptions but cannot remedy the more fundamental issue: Cox-derived p-values remain unreliable when the model is applied to data with nonlinear structures. This reliability problem persists regardless of whether the proportional hazards assumption is satisfied. Therefore, while we have provided the requested tests as a supplement, they do not alter our methodological conclusion that Spearman's approach is more appropriate for our nonlinear dataset.

To assess true associations between variables, we evaluated both consistency and dose-response relationships in our data. We implemented a leave-top1-out approach to examine the stability of feature

importance: first selecting top 10 features from the full dataset, then removing the highest-ranked feature to create a reduced dataset, re-selecting top 9 features from this reduced dataset, and comparing feature importance ranking orders between the original top 10 and new top 9 feature sets. This approach creates deliberate perturbation in the feature space, allowing us to observe how the removal of the dominant feature affects the overall ranking pattern—a critical test for consistency and dose-response relationships. Additionally, CV8 and CV5 feature sets were selected from the full dataset to provide complementary perspectives on feature importance.

For assessing true associations, we examined the consistency in ordered sets of features rather than merely identifying non-ordered collections of important variables. This distinction is crucial as ranking stability across different analytical conditions provides substantially stronger evidence of genuine relationships between predictors and outcomes. Table 1 presents these ordered sets of features across different cross-validation iterations, revealing remarkable consistency among key predictors even when the highest-ranked feature is systematically removed from consideration—further supporting the robustness of our identified associations.

This study set out to compare feature selection strategies for identifying informative air pollution constituents associated with mortality outcomes in an open benchmark dataset, using a downstream, held-out Random Forest regressor to isolate the effect of feature selection. Several key insights emerge that are relevant for environmental epidemiology and, by analogy, to the causal questions posed by Chen et al. regarding PM2.5 chemical constituents and lung cancer mortality.

First, nonparametric and unsupervised strategies outperformed Cox-based selection in this dataset. Spearman's rank correlation consistently yielded the highest cross-validated accuracy with both compact (top 5) and slightly expanded (top 8) feature sets, indicating that monotonic, assumption-light association screening can surface robust signals in settings characterized by skewed exposures, nonlinear relationships, and multicollinearity. Feature Agglomeration (FA) performed competitively with small feature sets, supporting the value of structure-preserving dimensionality reduction when predictors are highly correlated. Highly Variable Gene Selection (HVGS) was moderately effective but, as expected for a purely variability-driven method, it benefited less from outcome information than Spearman and was sensitive to how many features were retained. In contrast, Cox-based selection underperformed across configurations, consistent with a mismatch between proportional hazards and linearity assumptions and the underlying data-generating structure.

Second, these findings align with and extend the methodological message from Chen et al. While Chen et al. employed marginal structural Cox models with different weighting strategies to address time-varying confounding—a strength for causal inference—the performance gap we observe for Cox-based feature selection in an exploratory context underscores a broader point: model assumptions matter not only for unbiased estimation but also for which features are elevated as "important." When relationships are nonlinear, interactive, or time-varying, the semiparametric Cox framework can produce attenuated or unstable effect estimates that inadvertently down-rank genuinely informative exposures. Because many PM constituents co-vary and share sources, feature collinearity can further obscure individual effects in linear hazard models, amplifying instability in selected features.

Third, the superior performance of Spearman and FA highlights the practical value of a hybrid workflow that separates signal discovery from outcome modeling. By first characterizing structure (FA) and screening for robust monotonic associations (Spearman), analysts can reduce target-induced bias and mitigate overfitting to idiosyncratic patterns. Subsequent modeling—whether causal (e.g., marginal structural models, g-methods) or predictive (e.g., flexible machine learning)—can then be anchored in a more principled, lower-dimensional feature space. This approach is particularly pertinent when exploring which PM2.5 constituents warrant regulatory attention or mechanistic follow-up, where stability and interpretability of selected features matter as much as point estimates from any single model.

Fourth, our results reveal sensitivity to the number of selected features. Spearman's gains from 5 to 8 features suggest that modest expansion can capture additional complementary signal without overwhelming the model with noise, whereas FA's relative decline at 8 features indicates that cluster-representative selection may be most effective when kept compact. HVGS improved with a larger set, consistent with its broader, signal-agnostic filtering. These patterns argue for explicitly tuning the feature set size via nested cross-validation and for conducting ablation analyses to probe selection stability. We found that performance and rankings were more stable under Spearman and FA than under Cox-based selection, reinforcing the view that outcome-model misspecification can propagate into unstable feature importance.

Fifth, while the Random Forest regressor served as a neutral yardstick to compare feature sets, our findings should not be construed as an indictment of causal modeling. Rather, they suggest that causal analyses—especially those leveraging Cox-type models—benefit from front-end feature curation that respects correlation structure and nonlinear associations. In practice, one could: (a) use FA to define constituent clusters (e.g., secondary aerosols, traffic-related components, metals), (b) apply Spearman screening within clusters to identify representative constituents, (c) fit marginal structural models with flexible nuisance models for weights (e.g., boosted trees or ensemble learners), and (d) rigorously check proportional hazards, nonlinearity, and time-varying effects (e.g., spline terms, time-by-exposure interactions, additive hazards, accelerated failure time, or targeted maximum likelihood estimators). Such a pipeline bridges exploratory robustness with causal interpretability.

Limitations temper the generality of our conclusions. The benchmark dataset concerns COPD mortality, not lung cancer, and contains 21,028 complete observations across 19 unique features. Thus, observed performance gaps may differ in large-scale, high-dimensional settings, or when time-indexed exposures and confounders are available to build marginal structural models with finer temporal resolution. Our outcome was modeled via regression for pragmatic evaluation of feature sets, whereas time-to-event data in practice may demand survival-specific learners (e.g., random survival forests, gradient boosting for survival), which could shift relative performance. Additionally, Spearman's correlation, while robust and interpretable, captures only monotonic associations and can miss U-shaped or threshold effects; FA's clustering can be sensitive to distance metrics and linkage criteria; and HVGS depends on variance estimators and scaling choices. Finally, without true counterfactual exposure assignments, our comparisons assess predictive utility of features, not their causal effects.

Despite these caveats, the core implications are actionable for environmental health research.

- Diagnose and relax assumptions early. Before committing to proportional hazards, test for nonlinearity, non-proportionality, and interactions; consider flexible survival models or alternative causal estimators if diagnostics fail.
- Use unsupervised structure discovery to manage multicollinearity. Cluster correlated constituents and select representatives to reduce redundancy and enhance interpretability.
- Employ nonparametric association screening to prioritize stable signals. Spearman's rank correlations, with appropriate multiple-testing corrections, provide assumption-light evidence that travels well across resamples.
- Validate feature stability. Use cross-validated performance, permutation importance, and ablation to assess robustness; large swings in rankings suggest overreliance on a misspecified outcome model.
- Integrate exploratory and causal stages. After robust feature curation, fit causal models with flexible nuisance components (e.g.,

machine learning for propensity and censoring models), and report sensitivity to weighting strategies and functional forms.

In sum, our results support a pragmatic, hybrid workflow in which unsupervised and nonparametric tools precede and inform causal modeling. For questions like those posed by Chen et al.—identifying which PM2.5 constituents plausibly contribute to mortality—this strategy can yield more stable and credible feature sets, reduce susceptibility to model misspecification, and ultimately strengthen the transparency and reproducibility of inferred associations. Public benchmarks, even when imperfect proxies, are valuable testbeds for stress-testing methodological choices before application to proprietary cohorts.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The author has no permission to share data.

Code availability

Python code is publicly available at GitHub.

AI use

Not applicable.

Funding

This research has no fund.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Abd ElHafeez, S., D'Arrigo, G., Leonardis, D., et al., 2021. Methods to analyze time-to-event data: the cox regression analysis. Oxid. Med. Cell. Longev. 2021, 1302811. https://doi.org/10.1155/2021/1302811.

Austin, P.C., Giardiello, D., 2025. The impact of violation of the proportional hazards assumption on the calibration of the Cox proportional hazards model. Stat. Med. 44 (13–14), e70161. https://doi.org/10.1002/sim.70161.

Cai, W., Qi, Y., Zheng, L., et al., 2025. Comparison of random survival forest based-overall survival with deep learning and Cox proportional hazard models in HER-2-Positive HR-Negative breast cancer. Cancer Rep (Hoboken) 8 (7), e70262. https://doi.org/10.1002/cnr2.70262.

Chen, S., Ye, P., Wei, J., et al., 2025a. Causal links between long-term exposure to the chemical constituents of PM2.5 and lung cancer mortality in southern China. J. Hazard Mater. 495, 139096. https://doi.org/10.1016/j.jhazmat.2025.139096.

Chen, X., Liu, H., Men, J., You, J., 2025b. High-dimensional partially linear functional Cox models. Biometrics 81 (1), ujae164. https://doi.org/10.1093/biomtc/ujae164.

crossv, GitHub. Py with datasets of COPD mortality and air pollution analysis. https://github.com/y-takefuji/Cox. (Accessed 22 January 2026).

Jiang, N., Wu, Y., Li, C., 2024. Limitations of using COX proportional hazards model in cardiovascular research. Cardiovasc. Diabetol. 23 (1), 219. https://doi.org/10.1186/s12933-024-02302-2.

Jin, Y., Zhao, M., Su, T., et al., 2025. Comparing random survival forests and cox regression for nonresponders to neoadjuvant chemotherapy among patients with breast cancer: multicenter retrospective cohort study. J. Med. Internet Res. 27, e69864. https://doi.org/10.2196/69864.

Lin, T.A., McCaw, Z.R., Koong, A., et al., 2024. Proportional hazards violations in phase III cancer clinical trials: a potential source of trial misinterpretation. Clin. Cancer Res. 30 (20), 4791–4799. https://doi.org/10.1158/1078-0432.CCR-24-0566.

Qian, Z., Tian, L., Horiguchi, M., Uno, H., 2025. A novel stratified analysis method for testing and estimating overall treatment effects on time-to-event outcomes using average hazard with survival weight. Stat. Med. 44 (7), e70056. https://doi.org/10.1002/sim.70056.

Quantin, C., Asselain, B., Moreau, T., 1990. Le modèle de Cox: limites et extensions [The Cox model: limitations and extensions]. Rev. Epidemiol. Sante Publique 38 (4), 341–356.

Sandrock, C.E., Song, P.X.K., 2024. Limitation of site-stratified cox regression analysis in survival data: a cautionary tale of the PANAMO phase III randomized, controlled study in critically ill COVID-19 patients. Trials 25 (1), 822. https://doi.org/10.1186/s13063-024-08679-5.

Sharafi, M., Mohsenpour, M.A., Afrashteh, S., et al., 2024. Factors affecting the survival of prediabetic patients: comparison of Cox proportional hazards model and random survival forest method. BMC Med Inform Decis Mak 24 (1), 246. https://doi.org/10.1186/s12911-024-02648-3.

Xu, L., Jiang, S., Li, T., Xu, Y., 2024. Limitations of the cox proportional hazards model and alternative approaches in metachronous recurrence research. Gastric Cancer 27 (6), 1348–1349. https://doi.org/10.1007/s10120-024-01554-x.

Xu, L., Zhou, B., Xu, Y., 2025. Addressing biases: evaluating the Cox proportional hazards model and alternative approaches for major adverse cardiovascular events research. Eur Stroke J 10 (1), 298–299. https://doi.org/10.1177/23969873241286984.

Xue, J., Chen, Y., Xue, C., et al., 2024. Limitations of applying the COX proportional hazards model to glioma studies. J. Transl. Med. 22 (1), 1156. https://doi.org/10.1186/s12967-024-05942-w.