Can interpretability and accuracy coexist in cancer survival analysis?

Piyush Borole P.BOROLE@SMS.ED.AC.UK
Tongjie Wang TONGJIE.WANG@ED.AC.UK
Antonio Vergari AVERGARI@ED.AC.UK
Ajitha Rajan ARAJAN@ED.AC.UK
School of Informatics
University of Edinburgh
Edinburgh, UK

Abstract

Survival analysis refers to statistical procedures used to analyze data that focuses on the time until an event occurs, such as death in cancer patients. Traditionally, the linear Cox Proportional Hazards (CPH) model is widely used due to its inherent interpretability. CPH model help identify key disease-associated factors (through feature weights), providing insights into patient risk of death. However, their reliance on linear assumptions limits their ability to capture the complex, non-linear relationships present in real-world data. To overcome this, more advanced models, such as neural networks, have been introduced, offering significantly improved predictive accuracy. However, these gains come at the expense of interpretability, which is essential for clinical trust and practical application. To address the trade-off between predictive accuracy and interpretability in survival analysis, we propose ConSurv, a concept bottleneck model that maintains state-of-the-art performance while providing transparent and interpretable insights. Using gene expression and clinical data from breast cancer patients, ConSurv captures complex feature interactions and predicts patient risk. By offering clear, biologically meaningful explanations for each prediction, ConSurv attempts to build trust among clinicians and researchers in using the model for informed decision-making.

1. Introduction

Survival analysis is essential for estimating the time until events such as death, relapse, or recovery occur. It lays the groundwork for assessing disease severity and understanding the factors that influence patient outcomes (Clark et al. (2003); Bradburn et al. (2003)). Traditional methods like the **C**ox **P**roportional **H**azards (CPH) model, a linear approach, have been widely used due to their straightforward interpretability and robust ability to estimate hazard ratios (a type of relative risk) (Cox (1992); Bradburn et al. (2003)). However, with the advent of high-dimensional data such as **RNA seq**uencing (RNA-seq), these linear models face significant limitation in achieving high performance.

RNA expression data measure gene activity levels and play a critical role in survival research by revealing dysregulated genes associated with disease progression. The complex and non-linear relationships inherent in high-dimensional RNA-seq data challenge the CPH model's ability to accurately extract meaningful features. This limitation results in less

precise predictions compared to more advanced machine learning models that can capture these intricate patterns. Therefore, recently, deep learning approaches have been employed for survival analysis.

Accuracy-Interpretability trade-off

CPH, being a linear model, possesses inherent interpretability, allowing the importance of each input feature to be inferred from its feature weights. As illustrated in Figure 1, such a linear model offers high interpretability (on the x-axis) but low accuracy (due to low complexity, on the y-axis).

At the other extreme, emerging non-linear neural network models such as DeepSurv (Katzman et al. (2018)), offer improved accuracy by capturing the complexity of the data but provide very low interpretability. This lack of interpretability makes it challenging for clinicians and researchers to derive meaningful causal relationships between input features and risk predictions. This is illustrated in Figure 1, where the neural network ranks high on accuracy dimension but very low on interpretability dimension. The opacity of these models and the technical expertise required for their interpretation lead to a trust gap, preventing their adoption in clinical settings.

Tree-based models, such as XGBoost (Barnwal et al. (2022)), provide a partial solution by better balancing transparency and accuracy than deep learning models as seen in Figure 1. They offer transparency by using features as nodes within the trees, which could be individually examined, while also capturing non-linear relationships more effectively than linear models such as CPH. However, when applied to high-dimensional data such as RNA-seq, the trained trees can have vast number of decision nodes making them overwhelmingly complex. This growth in complexity inherently hinders interpretability.

As seen in the above examples of linear, tree-based, and neural network models, there is a trade-off between performance and interpretability, which is illustrated as accuracy-interpretability trade-off (or more appropriately complexity-interpretability) in Figure 1. This trade-off suggests that as model complexity increases, accuracy improves, but interpretability declines. An ideal model should balance both aspects to ensure usability and trustworthiness.

To balance the accuracy-interpretability trade-off in survival prediction, we propose a gray-box approach using concept bottleneck models (Koh et al. (2020)). These models consist of two parts: one predicts human-interpretable concepts, and the other—typically a linear model–uses these concepts for final predictions. We apply this approach to risk prediction on RNA-seq and clinical data for breast cancer and demonstrate performance comparable to the existing models. We also show that the key concepts can differentiate high- and low-risk patients demonstrating the usefulness of these concepts. Finally, we assess the stability of these concepts by evaluating their importance and robustness by examining the consistency in extracted concepts across multiple runs.

Code and data availability: The cancer dataset, code and complete list of packages used is available at: https://github.com/PRBorole/ConSurv.

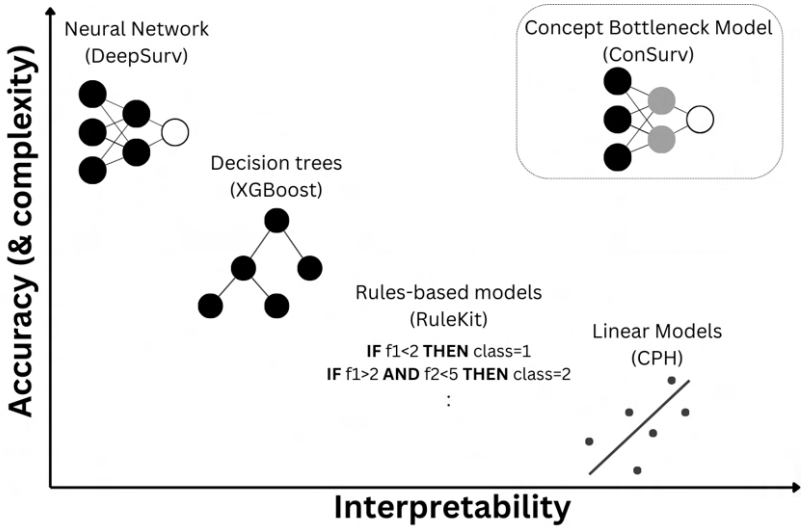

Figure 1: **Accuracy increases with model complexity, often at the expense of interpretability.** Simple models (e.g., linear or rule-based) are transparent but less accurate, while neural networks offer high accuracy but poor interpretability. Concept bottleneck models bridge this gap using intermediate, human-understandable concepts.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work highlights how concept-based models offer a practical middle ground between interpretable traditional models and high-performing black-box approaches in healthcare. By embedding clinically and biologically meaningful concepts (e.g., Estrogen Response) directly into the model architecture, ConSurv shows that *interpretability and accuracy can coexist*, a critical balance for clinical adoption. Importantly, even in the absence of ground-truth annotated concepts, which are often limited in biomedical domains, unsupervised concept extraction paired with domain-informed validation provides a scalable and effective alternative. While our focus is on survival analysis, this concept-based framework and validation strategy can be readily applied to other -omics datasets, (such as for predicting immunotherapy response or identifying molecular subtypes in cancer) especially as such data becomes increasingly prevalent. These contributions underscore the broader potential of gray-box models that deliver accurate predictions while fostering the transparency and trust essential for clinical AI adoption.

2. Related Work

Current research on interpretability of these neural networks for survival analysis focuses on post-hoc explanations which involve applying additional models to ascertain contribution of each feature (called explanations) to the predictions of these black-box models. For example, SurvLIME (Kovalev et al. (2020)) uses the **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations (LIME, Ribeiro et al. (2016)) framework, while SurvSHAP (Krzyziński et al.

(2023)) and AUTOSurv (Jiang et al. (2024)) apply **SH**apley **A**dditive ex**P**lanations (SHAP, Lundberg and Lee (2017)). Both LIME and SHAP are perturbation-based models that require multiple evaluation passes over the model to explain a single prediction. Furthermore, empirical evidence demonstrates that these models could fail to assign importance to relevant features and produce explanations that are unfaithful (Huang and Marques-Silva (2024)). These limitations make perturbation-based models not only computationally expensive but also, more importantly, unreliable (Rudin (2019); Bilodeau et al. (2024)). Other approaches, utilize DeepLIFT (Cho et al. (2023); Shrikumar et al. (2017)) and backpropagation (Yousefi et al. (2017)) to quantify the contribution of input features to the risk prediction. However, these gradient-based methods are highly sensitive to noise and nonlinearities, leading to explanations that are often unreliable and unfaithful to the models they aim to interpret (Ancona et al. (2019)). Relying on such explanations can lead to incorrect conclusions about the data, ultimately resulting in a trust deficit. The Cox-nnet model (Ching et al. (2018)), a single-layer perceptron, takes an interesting approach to interpretability by capturing biologically relevant functions (such as the p53 pathway) in its hidden nodes, which serve as surrogate features for predicting patient survival. However, their approach of quantifying the contribution of each input to every node of the network quickly becomes infeasible as deeper networks are required to improve accuracy. Our model, ConSurv, inherently embeds interpretability as a core component by aggregating genetic and clinical features into concepts, providing humanly understandable and biologically meaningful insights.

3. Methods

3.1. ConSurv

ConSurv is an interpretable concept bottleneck model that relies solely on concepts for its final predictions. Its interpretability stems from the use of interpretable concepts, while maintaining complete faithfulness by basing predictions exclusively on a linear combination of these concepts, as shown below:

$$\mathsf{LogRisk} = w_1 C_1 + w_2 C_2 + .. + w_m C_m \tag{1}$$

where $C_1..C_m$ are m interpretable concepts and $w_1..w_m$ their respective weights.

ConSurv has two layers: one hidden layer and one m-sized concept layer. The architecture of ConSurv is designed such that each concept receives input only from the features grouped for that concept (see Figure 2a). The concept layer output is normalized to values between 0–1 to ensure that the impact of each concept is influenced only by the static weight w for that concept.

To extract concepts, we rely on trained XGBoost trees and RuleKit rules, as illustrated in Figure 2b. Contrast set rule are a form of association rule learning which find patterns among the features that aim to distinguish between different groups. As illustrated in Figure 2, for each learned rule, the features are collected and grouped into concepts. On the other hands, tree based algorithms such as XGBoost are good at identifying features which are non-linearly related to each other and output. For XGBoost, concepts are extracted from each depth level of trained tree. In the illustration of Figure 2, the tree has two levels

nodes, *Node Pathology* feature at root and *Beta − hemoglobin* feature at level 1. There the extracted concept from root level is just the *Node Pathology* feature (Concept 1) and from level 1 is *Node Pathology* and *Beta − hemoglobin* feature (Concept 2).

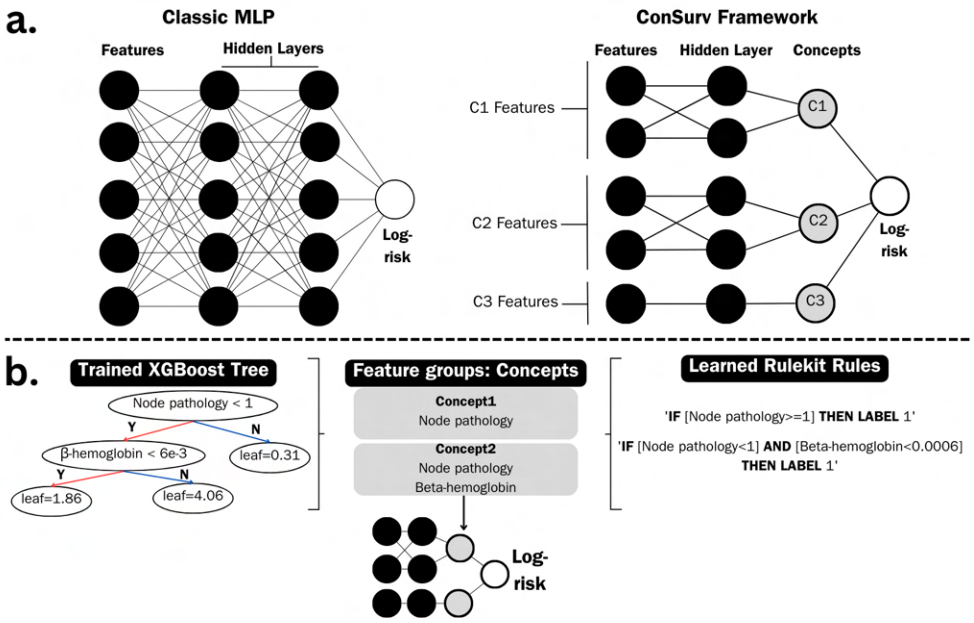

Figure 2: **ConSurv: A concept-bottleneck model for survival analysis.** A typical multi-layered perceptron consists of densely connected layers that capture non-linearities in the data, leading to high accuracy. However, this architecture makes the mapping from input to prediction difficult to interpret, hampering transparency. In contrast, ConSurv uses a concept layer, where each concept is based on a specific set of features, and the final LogRisk prediction is a linear combination of these concepts. Concepts are extracted using either a trained XGBoost model or RuleKit rules.

4. Cohort

4.1. Data

In this study, we utilized breast cancer RNA-seq (RNA profiling table) data and corresponding clinical information obtained from **T**he **C**ancer **G**enome **A**tlas (TCGA) through the **G**enomic **D**ata **C**ommons (GDC) portal. All data used in this study are de-identified and publicly available, eliminating the need for additional ethical approval. The study complies with the data usage policies of TCGA and GDC, which allow the use of the data for research purposes.

4.2. Data Extraction and Feature Choices

From TCGA, RNA profiling table, contains more than 60,000 gene expression values per patient were extracted. Given the limitations of physical memory and training time constraints, we opt for the variance threshold method to filter out genes that contain less information. We rank the index of dispersion for each gene and select the top 1000 highest-ranked genes. In addition to these 1000 genes, we added 8 clinical features (Such as Age, Race, Sex, etc.) associated with patients to the data set. Each patient entry was accompanied with survival time (t) and event indicator (e, which indicates if event, here death, occurred or not). The clinical features that are categorical (such as Race) were one hot encoded. Numerical features were min-max normalized before using for modeling. In total, there were 1209 patients with 1056 features in our final dataset.

5. Results

In the results section, we evaluate the models based on two key dimensions: performance in predicting LogRisk and ability in explaining model decision-making (i.e. interpretability). The existing models assessed in this study are CPH, XGBoost, and DeepSurv that are popular and widely used in survival analysis and covers the spectrum on accuracy-interpretability tradeoff of Figure 1. Performance is measured using the Concordance Index (or CI, see Appendix B.2).

Sections 5.1 and 5.2 present the performance evaluation of existing models and our proposed ConSurv framework. sections 5.3 to 5.5 focus on model interpretability. We begin by outlining the interpretability aspects of the linear CPH, the black-box DeepSurv (using SHAP), and the gray-box ConSurv model. This is followed by an exploration of the extracted concepts, assessing their alignment with biological insights, as well as their stability and robustness. The main sections focus on breast cancer, however, we obtain similar results for ovarian cancer, as presented in Appendix A. Table 1 provides a summary of the models used in this study and their interpretability.

Table 1: Survival Analysis Models And Their Interpretability

Model	**Type**	**White-/ Black-box**	**Interpretability**
CPH	Linear Model	White-box	Feature Weights
XGBoost	Non-Linear (Decision Trees)	White-box	Decision Tree, Post-hoc
DeepSurv	Non-Linear (Neural Network)	Black-box	Post-hoc
ConSurv	**Non-Linear (Concept Bottleneck Model)**	**Gray-Box**	**Learned Concepts**

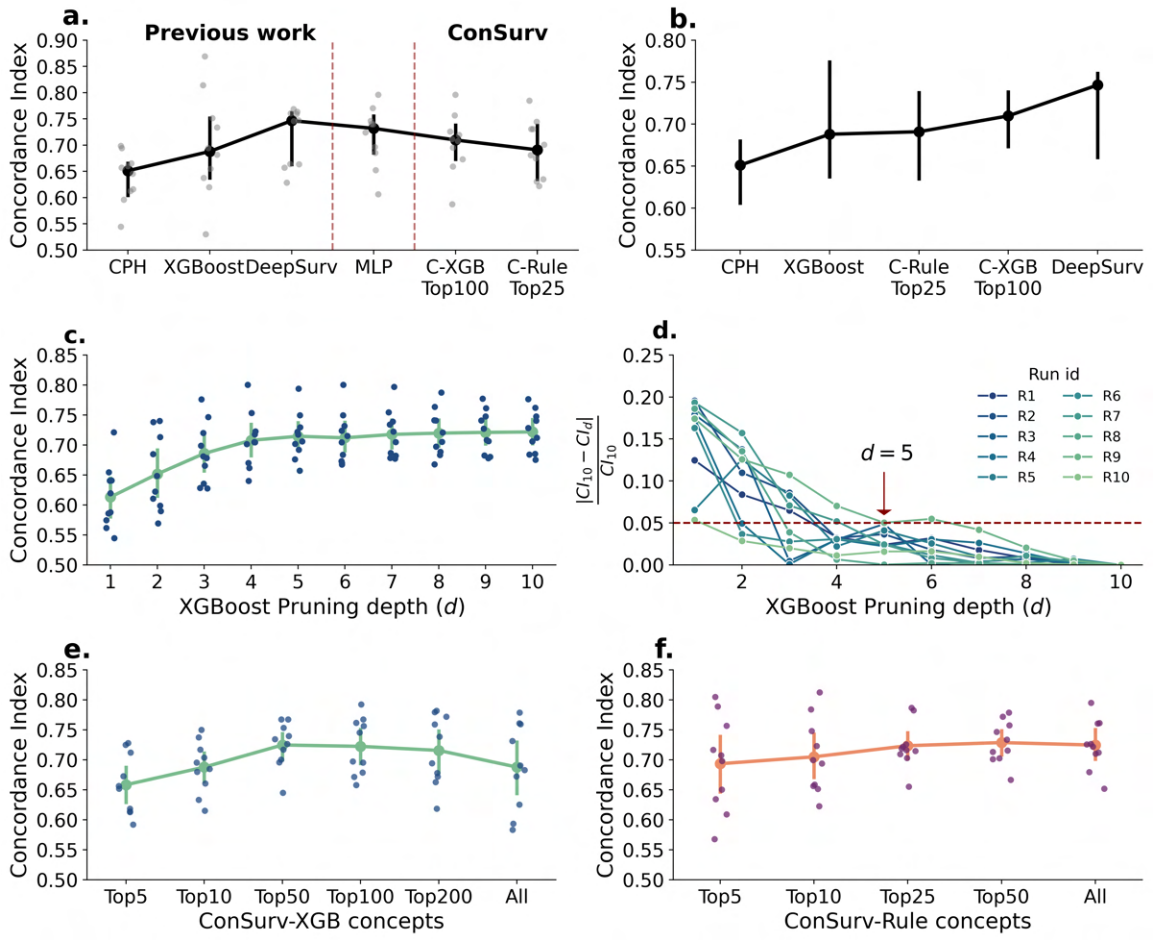

Figure 3: **ConSurv does not significantly trade off accuracy in predicting survival rates for interpretability.** a, b. CI increases as models transition from fully interpretable to black-box. Pruning XGBoost trees up to depth 5 from depth 10 retains performance within 5% (c, d). Top 100 for ConSurv-XGB and Top 25 for ConSurv-Rule concepts yields the highest performance (e,f)

5.1. Non-linear models demonstrate significantly better performance than linear model

In our study, we utilized widely used models - CPH, XGBoost, and DeepSurv, from the linear, tree-based, and neural network categories, respectively, as shown in Table 1 (and Figure 1). CPH is a linear model that provides association between the survival time of patients and the features. XGBoost is tree-based modeling approach used for survival regression that predicts survival time (Barnwal et al. (2022)). DeepSurv is a neural network that predicts LogRisk using Cox-loss function (See Appendix B.3, Katzman et al. (2018)). In addition to these, we trained a **M**ulti**L**ayer **P**erceptron (MLP) to establish baseline

parameters (such as network depth) for our concept bottleneck model. This base MLP was subsequently transformed to develop our ConSurv model as described in Section 3.1.

Linear CPH, exhibits the lowest median performance (median CI 0.65), while the black-box DeepSurv (median CI 0.75) achieves the highest median performance, with XGBoost (median CI 0.69) falling in between. The MLP demonstrates performance comparable to DeepSurv (median CI 0.73, p-value>0.05, see Appendix Table 6). Figures 3a,b plots the CI for the above described models.

For XGBoost, hyperparameter tuning revealed that the highest performance was achieved with a tree depth of 10. During each run, 330–390 trained trees were generated. XGBoost also exhibits the highest variance in performance, with the lowest CI being 0.53 and the highest CI reaching 0.87 as seen in Figure 3a. We pruned the trained XGBoost trees to depths ranging from 1 to 10 and found that pruning upto a depth of 5 resulted in a marginal loss in performance (within 5% of the model with a depth of 10), as shown in Figures 3c,d. An illustration of tree pruning is in Appendix Figure 9 which shows a tree with original depth 10 pruned upto depth 1. Across the runs, only 14 features (of 1056) were consistently used by XGBoost, as shown in Appendix Table 5. This low overlap of common features could explain the high variability in its performance. Nevertheless, the most commonly used features (predominantly genes) are known to be associated with breast cancer survival, and their references are provided in Appendix Table 5.

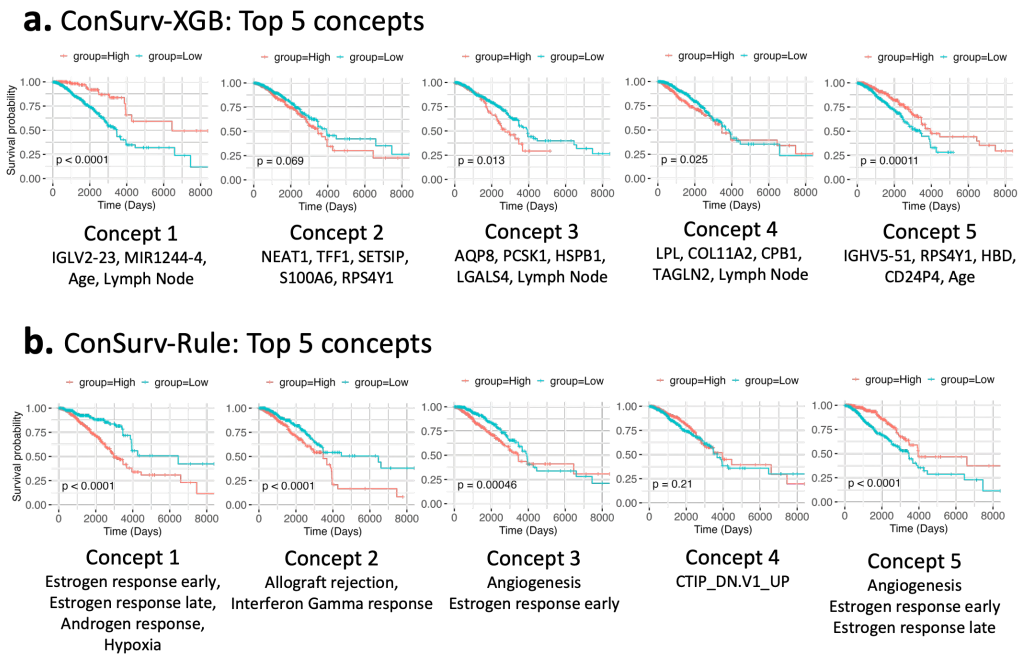

Figure 4: **The Kaplan-Meier (KM) plots for the top 5 concepts** from the ConSurv-XGB and ConSurv-Rule show significant separation between risk groups for all but one concept (concept 2 for ConSurv-XGB and concept 4 for ConSurv-Rule).

5.2. ConSurv maintains high performance with added interpretability

Our proposed model ConSurv, is a concept bottleneck model that predicts risk as described in Appendix Equation (3) with human interpretable concepts as an intermediate prediction. The architecture is based on the bilayered MLP model with the second layer modified to capture predetermined concepts as illustrated in Figure 2a. In our framework, concepts are defined as sets of grouped features extracted either from trained trees generated by XGBoost or from rules learned using the Contrast Set Rule-based model (RuleKit, see Appendix C), as described in Section 3.1. The two ConSurv models are named as ConSurv-XGB-all and ConSurv-Rule-all respectively. The '-all' indicates use of all concepts extracted.

ConSurv-XGB: Each run of XGBoost (pruned to depth 5) generated between 540 and 770 concepts. However, having such a large number of concepts can hinder the model's interpretability. To address this, we investigated reducing the number of concepts while maintaining performance. We first ranked the concepts from the trained ConSurv-XGB model based on their absolute weights. Then, we trained the model using only the top 5, 10, 50, 100, and 200 concepts. Appendix Figure 10 illustrates the ranked concepts in descending order of absolute weight across different runs. As shown in Figure 3e, the validation set CI improves with increasing number of concepts upto 100 but plateaus (and even slightly declines) beyond that, when using all concepts. Based on this, we consider ConSurv-XGB with the top 100 concepts (hereafter referred to simply as ConSurv-XGB or C-XGB Top100) to be the best model in our analysis, achieving a median test set CI of 0.71. This approach reduces the number of concepts by approximately 80–88%, significantly simplifying the model while maintaining performance.

ConSurv-Rule: Each run of RuleKit generated between 61 and 83 concepts, with an average of 25–30 features per concept. Similar to ConSurv-XGB, we evaluated models using the top 5, 10, 25, 50, and all available concepts. Appendix Figure 11 illustrates the ranked concepts in descending order of absolute weight across different runs. As shown in Figure 3f, validation set performance does not substantially improve beyond the top 25 concepts (median CI 0.69). We use the top 25 concept model's test set performance for comparison with existing models in Figure 3a (hereafter referred to simply as ConSurv-Rule or C-Rule Top25).

In terms of median test performance, ConSurv models' performance falls between XGBoost and DeepSurv. A Mann–Whitney U test confirms that ConSurv models, MLP, and DeepSurv perform significantly better than CPH. However, differences in performance among XGBoost, ConSurv models, MLP, and DeepSurv are not statistically significant (p-value > 0.05). The p-values for these comparisons are provided in Appendix Table 6.

We identified three key advantages of ConSurv. First, it achieves performance comparable to existing models. Second, it demonstrates stable performance across runs, as evidenced by a CI standard deviation of 0.06 for ConSurv-XGB, lower than the 0.1 observed for XGBoost, indicating consistency comparable to black-box models. Finally, as a gray-box model, ConSurv prioritizes interpretability through concepts as a core feature, ensuring that predictions rely exclusively on these concepts, thereby enhancing fidelity (or faithfulness). In the following sections, we explore the interpretability offered through concepts by ConSurv.

5.3. Concepts show clear high- vs. low-risk patient stratification

Interpretability of CPH: The CPH model has the lowest performance but provides inherent interpretability, with feature coefficients directly reflecting their importance. In Appendix Figure 12, we present the top 40 features ranked by their absolute value of CPH coefficients. With these coefficients one can recognize the contribution made by every feature to the final prediction. These coefficients are faithful to the model as they are directly used for risk prediction.

Interpretability of DeepSurv (SHAP limitations): In contrast to CPH, DeepSurv achieves the highest median performance among the evaluated models; however, interpreting its predictions requires post-hoc explanation techniques such as SHAP, which provides local explanations for individual patients. SHAP summary plot (Appendix Figure 13), which aggregate feature importance across all patients, indicate that *days_to_birth* (age) is the most important feature driving DeepSurv's predictions. However local explanations differ when we examined SHAP explanations for individual patients (Appendix Figure 14). For one patient, *days_to_birth* reduced the predicted risk (Appendix Figure 14a), while for another, it increased the risk (Appendix Figure 14b) and for a third patient (Appendix Figure 14c), the feature had no effect at all (SHAP score of 0.0). Notably, for the third patient, where SHAP suggested age was irrelevant, changing only the age resulted in changes in the predicted risk ($\Delta\mathsf{LogRisk} = 22$ and 15 for ages 25 and 80, respectively). This example highlights a key limitation of SHAP: even for the most important features, the explanations provided may not reliably reflect how the model truly uses them (SHAP only tries to explain prediction, not interpret the model). Similar variations were observed for other top feature *ajcc_pathology_t_T2*, which showed conflicting effects on risk across different patients. These findings underscore broader concerns regarding the faithfulness of post-hoc XAI techniques in accurately representing the inner workings of the models they are intended to explain, and there is growing research highlighting these limitations (Bilodeau et al. (2024); Rudin (2019); Slack et al. (2020)).

Interpretability of ConSurv: Original concept bottleneck models (Koh et al. (2020)) use predefined, human-interpretable concepts with ground truth for supervised concept training. However, availability of such concepts in many cases is rare. This is especially true in biomedical applications where AI is employed to gather insights from the data. In our case, clinicians not only want predictions but also features important for those predictions. In such scenarios, unsupervised extraction of concepts is a possible strategy where the extracted concepts are later ratified with domain knowledge based on the features they capture. The work by Wu et al. (2022) demonstrates a time-series concept bottleneck model using this strategy. We use a similar approach where we first extract concepts by grouping features using either XGBoost or RuleKit (See Figure 2b, details in Section 3.1). Appendix Tables 7 to 12 provides the top five concepts for both ConSurv models.

We evaluate the effectiveness of each of the top five concepts from both models in distinguishing high- and low-risk groups using Kaplan-Meier survival plots. We obtained the output for all patients for each of the top five concepts using the ConSurv-XGB and ConSurv-Rule models. These outputs were then used to generate the KM plots presented in Figure 4 (c for ConSurv-XGB and d for ConSurv-Rule). For the ConSurv-XGB model, all concepts except concept 2 showed significant stratification (i.e., $p < 0.05$) between high-

and low-risk patients. Similarly, for the ConSurv-Rule model, all concepts except concept 4 demonstrated significant differentiation (i.e., $p < 0.05$) between patient groups. This suggests that the top five most influential concepts in risk prediction from both models can effectively distinguish patients.

Our framework addresses the limitations of previous work through interpretable concepts that group features together, offering insight into potential interactions. Furthermore, because the final prediction relies exclusively on these concepts, they present faithfulness in the predictions. Additionally, it provides concept importance through the learned weights assigned to each concept. In the next section, we analyze and interpret these top five concepts in the context of their biological relevance.

5.4. Concepts capture biological processes

5.4.1. Interpretation of top 5 ConSurv-XGB concepts

Concept 1: The top-ranked concept in ConSurv-XGB, which exhibits the most significant stratification between high- and low-risk groups. This concept comprises the features: age, lymph, IGLV2-23, and MIR1244-4. Age is a key factor in survival outcomes, with women under 40 and over 80 showing poorer prognosis (Brandt et al. (2015)). Lymph node pathology is also crucial, as one pathway for breast cancer metastasis is through the lymphatic system and lymph nodes (Nathanson et al. (2022)). This is especially significant for women under 40 with axillary lymph node-negative breast cancer shown to have poor prognosis (Brandt et al. (2015)). Older patients with a high lymph node ratio are shown to have a threefold increased risk of breast cancer death (Wildiers et al. (2009); Vinh-Hung et al. (2010)). The IGLV family (here, IGLV2-23) is linked to older age at diagnosis and a distinct stromal microenvironment in breast cancer Brouwers et al. (2017). This concept connects age, lymph node pathology, and the IGLV gene. While we couldn't find details of mechanism for MIR1244-2, there is increasing evidence of microRNAs' role in breast cancer malignancy (Muñoz et al. (2023)).

Concept 2: comprises of NEAT1, TFF1, S100A6, RPS4Y1 and SETSIP genes. NEAT1 (Shin et al. (2019)) promotes breast cancer, while TFF1 (Yi et al. (2020)) and S100A6 (Qi et al. (2023)) suppress it. All three interact with or affect estrogen expression (Knutsen et al. (2022); Yi et al. (2020); Desai et al. (2005)). RPS4Y1 (Li et al. (2024)), a male breast cancer marker, may serve as a proxy for estrogen status. SETSIP, while not directly linked, is associated with angiogenesis (Margariti et al. (2012)), potentially supporting tumor growth. This concept reflects estrogen response in breast cancer.

Concept 3: includes four genes—AQP8, PCSK1, HSPB1, and LGALS4—along with one clinical feature, lymph node pathology. AQP8 is highly expressed in basal and luminal B breast cancer types, which have low estrogen receptor (ER) levels (Zhu et al. (2019)). PCSK1 is upregulated in breast cancer, promoting tumor progression, estrogen dependency, and anti-estrogen resistance in cell lines (Jaaks and Bernasconi (2017)). HSPB1 is linked to metastasis via epithelial-to-mesenchymal transition and is correlated with lymph node status and estrogen receptors (Huo et al. (2023)). LGALS4 high expression is a good prognostic factor for LN-negative patients (Grosset et al. (2016)). This concept highlights estrogen dependency (via AQP8, PCSK1, and HSPB1) and lymph node involvement (via lymph node pathology, HSPB1, and LGALS4).

Concept 4: includes features COL11A2, TAGLN2, LPL, CPB1 and lymph node pathology. COL11A2 (Luo et al. (2022)) and TAGLN2 (Xu et al. (2010); Meng et al. (2017)) are linked to lymph node pathology. LPL promotes tumor growth by altering the microenvironment through lipid hydrolysis (Bavis et al. (2023)), while all three are potential therapeutic targets. CPB1 down-regulates tumor suppressors like SFRP1 and OS9 (Kothari et al. (2021)) and is up-regulated in lymph node-positive patients (Bouchal et al. (2015)). This concept highlights features associated with lymph node involvement and potential therapeutic targets.

Concept 5: comprises of HBD, CD24, IGHV5-51, RPS4Y1 and age. The loss of the immunomodulatory HBD (Pandurangi et al. (2024)) and overexpression of CD24 (Kristiansen et al. (2003); Huang et al. (2024)) promote metastasis by enabling escape from cell death. The IGHV family (e.g., IGHV5-51) is linked to vacuolation and degeneration in mouse breast tumorigenesis (Ganaie et al. (2020)). Together, these features suggest immune dysregulation, promoting cell death avoidance and tumor proliferation. However, the role of age and RPS4Y1 in conjunction with these genes requires further exploration.

We find ConSurv-XGB concepts are smaller in size (only 1–5 features per concept), with each capturing a biologically meaningful property. Further, understanding relationships between features within each concept often requires domain knowledge as seen for the five concepts discussed above.

5.4.2. Interpretation of top 5 ConSurv-Rule concepts

The concepts extracted from RuleKit tend to be larger (25–30 features per concept), leading us to hypothesize that they capture broader biological processes. To investigate this, we performed **O**ver**R**epresentation **A**nalysis (ORA) (See Appendix D.2) to identify associated processes, with results listed in Table 2. We used the Hallmark and C6 gene set databases in MSigD to determine relevant processes. Hallmark gene sets represent well-defined biological states with coherent expression patterns and were our primary reference. When no significant process were identified in Hallmark, we turned to C6. We found that all top 5 concepts in RuleKit can be associated with broader biological processes. Concept 1 primarily captures signaling pathways but also identifies hypoxia, a key feature of the tumor microenvironment that promotes angiogenesis in breast cancer (Elayat and Selim (2024)). Concept 2 is associated with immune-related processes, as indicated by the Hallmark allograft rejection and interferon-gamma response gene sets, both critical in breast cancer progression (Oshi et al. (2021, 2022)). Concept 4 lacks Hallmark features but captures the C6 oncogenic signature CTIP_DN.V1_UP, linked to early-onset breast cancer (Zarrizi et al. (2020)).

Two key characteristics of RuleKit concepts: 1. Certain gene sets, such as estrogen response and angiogenesis, appears repeatedly across concepts (both processes are crucial for breast cancer survival, Oshi et al. (2020); Elayat and Selim (2024)); 2. they often encompass multiple biological processes suggesting that RuleKit identifies high-level biological pattern.

Both ConSurv models provide biologically meaningful concepts. While ConSurv-XGB requires domain knowledge to interpret feature associations within each concept, ConSurv-Rule concepts can be more easily linked to well-established biological processes. Unlike conventional methods that rely on individual feature weights or costly and often unreliable

Table 2: Biological processes of top 5 ConSurv-Rule concepts

Concept	Biological Process
Concept 1	HALLMARK_ESTROGEN_RESPONSE_EARLY
	HALLMARK_ESTROGEN_RESPONSE_LATE
	HALLMARK_ANDROGEN_RESPONSE
	HALLMARK_HYPOXIA
Concept 2	HALLMARK_ALLOGRAFT_REJECTION
	HALLMARK_INTERFERON_GAMMA_RESPONSE
Concept 3	HALLMARK_ANGIOGENESIS
	HALLMARK_ESTROGEN_RESPONSE_EARLY
Concept 4	CTIP_DN.V1_UP
Concept 5	HALLMARK_ESTROGEN_RESPONSE_EARLY
	HALLMARK_ESTROGEN_RESPONSE_LATE
	HALLMARK_ANGIOGENESIS

post-hoc explanations, ConSurv models offer direct biological insights through their inherently interpretable design. In the next section, we analyze `Stability` and `Robustness` of the concepts from both models.

5.5. Stability and Robustness

In this section, we analyze two important aspects of our model, namely – `Stability`: Whether concept importance remains same?; `Robustness`: Do concepts vary across different data splits. Essentially, `Stability` evaluates ConSurv network, `Robustness` is a metric of evaluation of the concept extraction process.

Stability: We examined whether training the model using only the top-ranking concepts impacts their rankings. A model that preserves most of the concept rankings would suggest that these concepts and their importance remain stable. We compared the ranking of 100 concepts in ConSurv-XGB model with original rankings from ConSurv-XGB with all concepts. We used Kendall's Tau to assess the similarity of rankings and found that it ranged between -0.08 to 0.32 (plotted in Figure 5a, Appendix Figure 8d) which indicates there is little to no correlations between the two rankings. Similarly, we compared the ranking of the 25 concepts in ConSurv-Rule model with original rankings from ConSurv-Rule with all concepts. We found that the Kendall's Tau value ranged between -0.19 to 0.56 (plotted in Figure 5b, Appendix Figure 8d) which indicates there is little correlations (better than ConSurv-XGB) between the two rankings. These results, which show a high mismatch between the original and retrained rankings for both model concepts, indicate that the concepts and their importance are not `stable`.

Robustness: The concept extraction process can be considered robust if the concepts extracted across multiple runs exhibit a high degree of similarity. We evaluated the similarity of concepts across runs in both ConSurv models. Concepts represent sets of features, with sizes ranging from 1–5 features for ConSurv-XGB and averaging 25–30 features for ConSurv-Rule. To measure similarity, we used the commonly applied Jaccard Similarity Index (SI). However, as Jaccard SI is influenced by set size, we also used the Cosine SI, which focuses on overall similarity between sets. Figures 5c,d shows the distributions of

Jaccard and Cosine SI across the 45 run pairs (10 runs yields 45 pairs), comparing every concept from one run to every concept in another run. For each pair, comparing 100 concepts in ConSurv-XGB results in 4,950 comparisons and 300 comparisons for the 25 concepts in ConSurv-Rule. ConSurv-Rule concepts were noticeably more consistent across runs compared to those from XGBoost, as indicated by higher Jaccard and Cosine similarity indices. This consistency is further supported by two observations: First, a trained RuleKit model utilizes more features per run than a trained XGBoost model (43%–53% vs. 35%–50%, respectively), allowing for more potential overlap across runs. Second, 44 features were consistently present across RuleKit runs, compared to only 14 features in XGBoost runs. However, despite this relative difference, the overall similarity indices for both remained quite low, indicating limited overlap in the concepts across runs.

In summary, ConSurv-XGB concepts are noticeably smaller than ConSurv-Rule concepts. Similarity in concepts across runs is limited for both models. However, in contrast to ConSurv-XGB, ConSurv-Rule concepts are more consistent across runs. Additionally, concepts from both model appear not stable, as their rankings change after retraining.

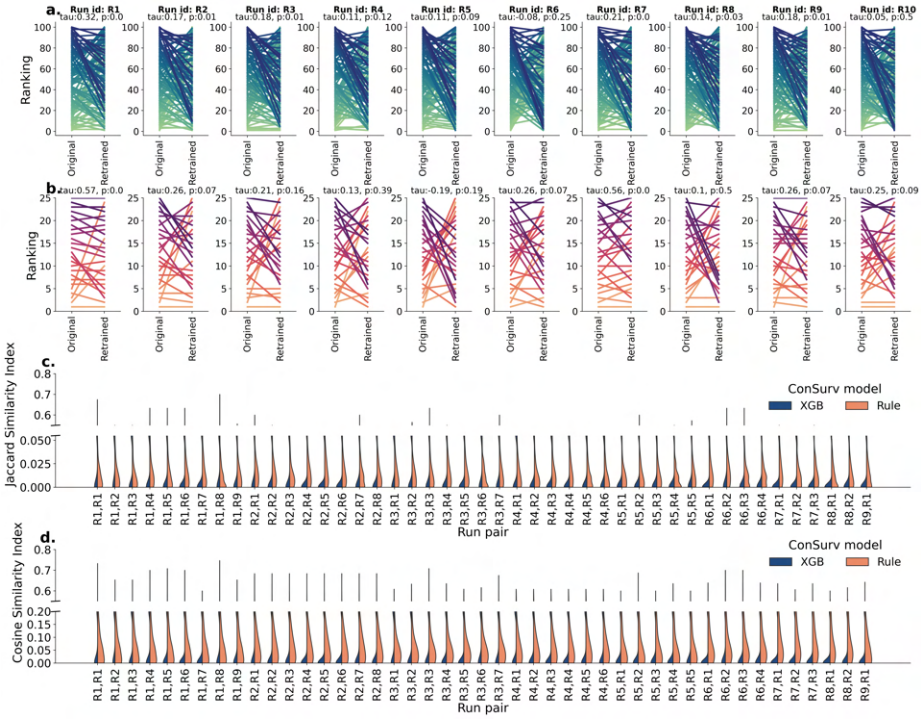

Figure 5: **ConSurv-Rule concepts are more stable and consistent across runs than ConSurv-XGB.** Concept ranking `stability`, assessed via Kendall's Tau, was higher for ConSurv-Rule (-0.19 to 0.56) than ConSurv-XGB (–0.8 to 0.32). Both Jaccard and Cosine Similarity indices show limited overlap across runs, though ConSurv-Rule concepts are more consistent.

6. Discussion

Our study aims to improve interpretability in survival analysis while balancing the trade-off with performance. Through an evaluation of existing models and our proposed ConSurv framework, we demonstrated that incorporating interpretable concepts enhances model transparency without compromising performance.

Trade-off between interpretability and accuracy: The CPH, a fully white-box approach, offers inherent interpretability via feature coefficients but suffers from the lowest median performance. Conversely, black-box models such as DeepSurv achieve the highest median performance but require unreliable post-hoc explanations. ConSurv, a gray-box model, successfully balances these concerns. By leveraging interpretable concepts from ConSurv-XGB and ConSurv-Rule, we achieve performance similar to high performing models while providing human-interpretable insights. Importantly, ConSurv-XGB demonstrates lower CI variance than XGBoost, indicating increased stability of the performance.

Concepts enable significant stratification among patients: A key advantage of our framework lies in its ability to extract and leverage meaningful concepts in survival prediction. Unlike most post-hoc methods that provide individual feature importance without capturing interactions, ConSurv integrates concepts directly into the model. This allows a more interpretable and faithful predictions with biologically relevant representation of survival risk factors. This is evident in our analysis of concepts and their biological significance. Additionally, KM survival plots (Figure 4) show that top ConSurv-XGB and ConSurv-Rule provides significant survival stratification between high- and low-risk patients.

Biological relevance of concepts: To assess the interpretability and biological relevance of our learned concepts, we performed ORA to visualize gene interaction networks. The extracted concepts align with known biological processes relevant to cancer survival, such as hypoxia, angiogenesis, etc. We apply very stringent condition that only Hallmark (preferred) and C6 databases were used to ensure only most relevant processes are obtained. However, this can be left at the discretion of the end user to set the criteria.

Stability and reproducibility of extracted concepts: An important consideration in modeling approach is the consistency of learned concepts across different runs. Our analysis of concept `robustness` (Figure 5) reveals that ConSurv-Rule concepts exhibit greater consistency across runs compared to ConSurv-XGB concepts. However, despite this advantage, overall concept similarity remains low, indicating variability in the specific features grouped within each concept. Additionally, the stability of concepts for both models was low. More research is needed to develop strategies for extracting concepts more effectively, ensuring greater consistency across different runs, thereby allowing the ConSurv framework to be more readily used.

Limitations: Despite the advantages of ConSurv, some limitations warrant further investigation. First, the variability in concept extraction across runs suggests a need for more robust feature selection or concept aggregation strategies. One potential approach is to integrate ensemble learning techniques that uniformize concept extraction across multiple runs. Additionally, while ConSurv-Rule concepts are relatively robust, they include many

features per concept which capture multiple processes, and exhibit redundancy across concepts. Furthermore, when generating rules, treating censored data as a lower bound may lead to the loss of valuable information. Therefore, as a future step, a more robust temporal discretization strategy should be explored. ConSurv-XGB concepts are small, but interpreting their biological meaning requires extensive domain knowledge. Our current work focuses on tabular data, however, recently survival analysis is done incorporating multiple modalities, such as images, which could also be integrated into a concept-based framework easily.

Conclusion: Our study introduces ConSurv, a gray-box survival analysis framework that balances interpretability and accuracy by leveraging concept-based learning. Through comparisons with existing models, we demonstrate that ConSurv achieves high predictive performance while maintaining transparency in its decision-making with biologically relevant concepts. This unique ability positions ConSurv as a valuable tool for clinical decision support, enabling both accurate risk prediction and actionable insights. Future research will focus on robust concept extraction and exploring broader applications in precision medicine.

Acknowledgments

This work and authors P.B., A.R., are supported by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101017453. A.V. is supported by the "UNREAL: Unified Reasoning Layer for Trustworthy ML" project (EP/Y023838/1) selected by the ERC and funded by UKRI EPSRC. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/). The authors also thank G-Research for its financial support and the opportunity to present this research.

References

J Adnane, P Gaudray, CA Dionne, G Crumley, M Jaye, J Schlessinger, P Jeanteur, D Birnbaum, and C Theillet. Bek and flg, two receptors to members of the fgf family, are amplified in subsets of human breast cancers. *Oncogene*, 6(4):659–663, 1991.

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Gradient-based attribution methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 169–191, 2019.

Gayatri Arun and David L Spector. Malat1 long non-coding rna and breast cancer. *RNA biology*, 16(6):860–863, 2019.

Avinash Barnwal, Hyunsu Cho, and Toby Hocking. Survival regression with accelerated failure time model in xgboost. *Journal of Computational and Graphical Statistics*, 31(4): 1292–1302, 2022.

Makayla M Bavis, Allison M Nicholas, Alexandria J Tobin, Sherri L Christian, and Robert J Brown. The breast cancer microenvironment and lipoprotein lipase: Another negative notch for a beneficial enzyme? *FEBS Open Bio*, 13(4):586–596, 2023.

Stephen D Bay and Michael J Pazzani. Detecting group differences: Mining contrast sets. *Data mining and knowledge discovery*, 5:213–246, 2001.

Marc B Bechmann, Andreas V Brydholm, Victoria L Codony, Jiyoung Kim, and René Villadsen. Heterogeneity of ceacam5 in breast cancer. *Oncotarget*, 11(43):3886, 2020.

Alakesh Bera, Madhan Subramanian, John Karaian, Michael Eklund, Surya Radhakrishnan, Nahbuma Gana, Stephen Rothwell, Harvey Pollard, Hai Hu, Craig D Shriver, et al. Functional role of vitronectin in breast cancer. *PLoS One*, 15(11):e0242141, 2020.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024.

Evan P Booy, Ewan KS McRae, Amit Koul, Francis Lin, and Sean A McKenna. The long non-coding rna bc200 (bcyrn1) is critical for cancer cell survival and proliferation. *Molecular cancer*, 16:1–15, 2017.

Pavel Bouchal, Monika Dvořáková, Theodoros Roumeliotis, Zbyněk Bortlíček, Ivana Ihnatová, Iva Procházková, Jenny TC Ho, Josef Maryáš, Hana Imrichová, Eva Budinská, et al. Combined proteomics and transcriptomics identifies carboxypeptidase b1 and nuclear factor κb (nf-κb) associated proteins as putative biomarkers of metastasis in low grade breast cancer. *Molecular & Cellular Proteomics*, 14(7):1814–1830, 2015.

Mike J Bradburn, Taane G Clark, Sharon B Love, and Douglas Graham Altman. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3):431–436, 2003.

Jasmine Brandt, Jens Peter Garne, Ingrid Tengrup, and Jonas Manjer. Age at diagnosis in relation to survival following breast cancer: a cohort study. *World journal of surgical oncology*, 13:1–11, 2015.

Heather Ann Brauer, Monica D'Arcy, Tanya E Libby, Henry J Thompson, Yutaka Y Yasui, Nobuyuki Hamajima, Christopher I Li, Melissa A Troester, and Paul D Lampe. Dermcidin expression is associated with disease progression and survival among breast cancer patients. *Breast cancer research and treatment*, 144:299–306, 2014.

Barbara Brouwers, Debora Fumagalli, Sylvain Brohee, Sigrid Hatse, Olivier Govaere, Giuseppe Floris, Kathleen Van den Eynde, Yacine Bareche, Patrick Schöffski, Ann Smeets, et al. The footprint of the ageing stroma in older patients with breast cancer. *Breast Cancer Research*, 19:1–14, 2017.

Demet Candas, Chung-Ling Lu, Ming Fan, Frank YS Chuang, Colleen Sweeney, Alexander D Borowsky, and Jian Jian Li. Mitochondrial mkp1 is a target for therapy-resistant her2-positive breast cancer cells. *Cancer research*, 74(24):7498–7509, 2014.

Shih-Hsuan Chan, Hsuan-Jung Tseng, and Lu-Hai Wang. Cd24a knockout transforms the tumor microenvironment from cold to hot by promoting tumor-killing immune cell infiltration in a murine triple-negative breast cancer model. *bioRxiv*, pages 2024–07, 2024.

Allan S Chen, Hongliang Liu, Yufeng Wu, Sheng Luo, Edward F Patz Jr, Carolyn Glass, Li Su, Mulong Du, David C Christiani, and Qingyi Wei. Genetic variants in ddo and pex5l in peroxisome-related pathways predict non-small cell lung cancer survival. *Molecular carcinogenesis*, 61(7):619–628, 2022.

Weiwei Chen, Xia Li, Youqin Jiang, Daguang Ni, Longfei Yang, Jixiang Wu, Mingcheng Gao, Jin Wang, Jianxiang Song, and Wenyu Shi. Pancancer analysis of the correlations of hs6st2 with prognosis, tumor immunity, and drug resistance. *Scientific Reports*, 13(1): 19209, 2023.

Venugopalan Cheriyath, Jaspreet Kaur, Anne Davenport, Ashjan Khalel, Nobel Chowdhury, and Lalitha Gaddipati. G1p3 (ifi6), a mitochondrial localised antiapoptotic protein, promotes metastatic potential of breast cancer cells through mtros. *British journal of cancer*, 119(1):52–64, 2018.

Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.

Hyun Jae Cho, Mia Shu, Stefan Bekiranov, Chongzhi Zang, and Aidong Zhang. Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment. *Bioinformatics*, 39(4):btad113, 2023.

Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

David R Cox. Regression models and life-tables. breakthroughs in statistics. *Stat. Soc*, 372: 527–541, 1992.

Yingnan Cui, Yan Jiao, Keren Wang, Miao He, and Zhaoying Yang. A new prognostic factor of breast cancer: High carboxyl ester lipase expression related to poor survival. *Cancer genetics*, 239:54–61, 2019.

Kartiki V Desai, JL Simmons, A Fargiano, G Van Den Eynden, PB Vermeulen, LY Dirix, M Merino, and JE Green. S100a6 as a biomarker in human breast cancer. In *Proc Amer Assoc Cancer Res*, volume 46, page 448, 2005.

Ghada Elayat and Abdel Selim. Angiogenesis in breast cancer: insights and innovations. *Clinical and Experimental Medicine*, 24(1):178, 2024.

Ethan D Emberley, Leigh C Murphy, and Peter H Watson. S100a7 and the progression of breast cancer. *Breast Cancer Research*, 6:1–7, 2004.

Juliana Oliveira Fernandes, Cassio Cardoso-Filho, Maria Beatriz Kraft, Amanda Sacilotto Detoni, Barbara Narciso Duarte, Julia Yoriko Shinzato, and Diama Bhadra Vale. Differences in breast cancer survival and stage by age in off-target screening groups: a population-based retrospective study. *AJOG Global Reports*, 3(2):100208, 2023.

Ishfaq Ahmad Ganaie, Md Zubbair Malik, Samar Husain Naqvi, Swatantra Kumar Jain, and Saima Wajid. Differential levels of alpha-1-inhibitor iii, immunoglobulin heavy chain variable region, and hypertrophied skeletal muscle protein gtf3 in rat mammary tumorigenesis. *Biochimie*, 174:57–68, 2020.

Debolina Ganguly, Marcel O Schmidt, Morgan Coleman, Tuong-Vi Cindy Ngo, Noah Sorrelle, Adrian TA Dominguez, Gilbert Z Murimwa, Jason E Toombs, Cheryl Lewis, Yisheng V Fang, et al. Pleiotrophin drives a prometastatic immune niche in breast cancer. *Journal of Experimental Medicine*, 220(5):e20220610, 2023.

Isabelle Grootes, Gordon C Wishart, and Paul David Peter Pharoah. An updated predict breast cancer prognostic model including the benefits and harms of radiotherapy. *NPJ Breast Cancer*, 10(1):6, 2024.

Andrée-Anne Grosset, Marilyne Labrie, Maria Claudia Vladoiu, Einas M Yousef, Louis Gaboury, and Yves St-Pierre. Galectin signatuares contribute to the heterogeneity of breast cancer and provide new prognostic information and therapeutic targets. *Oncotarget*, 7(14):18183, 2016.

Xiao-Li Gu, Zhou-Luo Ou, Feng-Juan Lin, Xiao-Li Yang, Jian-Min Luo, Zhen-Zhou Shen, and Zhi-Ming Shao. Expression of cxcl14 and its anticancer role in breast cancer. *Breast cancer research and treatment*, 135:725–735, 2012.

Adam Gudyś, Marek Sikora, and Łukasz Wróbel. Rulekit: A comprehensive suite for rule-based learning. *Knowledge-Based Systems*, 194:105480, 2020.

Adam Gudyś, Marek Sikora, and Łukasz Wróbel. Separate and conquer heuristic allows robust mining of contrast sets in classification, regression, and survival data. *Expert Systems with Applications*, page 123376, 2024.

Sonja Hänzelmann, Robert Castelo, and Justin Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, 14:1–15, 2013.

Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

Shiming Huang, Xiaobo Zhang, Yingtian Wei, and Yueyong Xiao. Checkpoint cd24 function on tumor and immunotherapy. *Frontiers in Immunology*, 15, 2024. URL https://api.semanticscholar.org/CorpusID:268212342.

Xuanxiang Huang and Joao Marques-Silva. On the failings of shapley values for explainability. *International Journal of Approximate Reasoning*, page 109112, 2024.

Qin Huo, Juan Wang, and Ni Xie. High hspb1 expression predicts poor clinical outcomes and correlates with breast cancer metastasis. *BMC cancer*, 23(1):501, 2023.

Patricia Jaaks and Michele Bernasconi. The proprotein convertase furin in tumour progression. *International journal of cancer*, 141(4):654–663, 2017.

Lindong Jiang, Chao Xu, Yuntong Bai, Anqi Liu, Yun Gong, Yu-Ping Wang, and Hong-Wen Deng. Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. *NPJ precision oncology*, 8(1):4, 2024.

Anbarasu Kannan, Julie V Philley, Kate L Hertweck, Harrison Ndetan, Karan P Singh, Subramaniam Sivakumar, Robert B Wells, Ratna K Vadlamudi, and Santanu Dasgupta. Cancer testis antigen promotes triple negative breast cancer metastasis and is traceable in the circulating extracellular vesicles. *Scientific reports*, 9(1):11632, 2019.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18:1–12, 2018.

Jongchan Kim, Hai-Long Piao, Beom-Jun Kim, Fan Yao, Zhenbo Han, Yumeng Wang, Zhenna Xiao, Ashley N Siverly, Sarah E Lawhon, Baochau N Ton, et al. Long noncoding rna malat1 suppresses breast cancer metastasis. *Nature genetics*, 50(12):1705–1715, 2018.

Noriko Kimura, Ryuichi Yoshida, Shin-ichiro Shiraishi, Monika Pilichowska, and Noriaki Ohuchi. Chromogranin a and chromogranin b in noninvasive and invasive breast carcinoma. *Endocrine pathology*, 13:117–122, 2002.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Erik Knutsen, Adrian L Harris, and Maria Perander. Expression and functions of long non-coding rna neat1 and isoforms in breast cancer. *British journal of cancer*, 126(4): 551–561, 2022.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

Charu Kothari, Alisson Clemenceau, Geneviève Ouellette, Kaoutar Ennour-Idrissi, Annick Michaud, Caroline Diorio, and Francine Durocher. Is carboxypeptidase b1 a prognostic marker for ductal carcinoma in situ? *Cancers*, 13(7):1726, 2021.

Maxim S Kovalev, Lev V Utkin, and Ernest M Kasimov. Survlime: A method for explaining machine learning survival models. *Knowledge-Based Systems*, 203:106164, 2020.

Glen Kristiansen, Klaus-Jürgen Winzer, Empar Mayordomo, Joachim Bellach, Karsten Schlüns, Carsten Denkert, Edgar Dahl, Christian Pilarsky, Peter Altevogt, Hans Guski, et al. Cd24 expression is a new prognostic marker in breast cancer. *Clinical cancer research*, 9(13):4906–4913, 2003.

Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, and Przemysław Biecek. Survshap (t): time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262:110234, 2023.

Soojung Lee, Nicolai J Toft, Trine V Axelsen, Maria Sofia Espejo, Tina M Pedersen, Marco Mele, Helene L Pedersen, Eva Balling, Tonje Johansen, Mark Burton, et al. Carbonic anhydrases reduce the acidity of the tumor microenvironment, promote immune infiltration, decelerate tumor growth, and improve survival in erbb2/her2-enriched breast cancer. *Breast Cancer Research*, 25(1):46, 2023.

Travis Leung, Ramkumar Rajendran, Subir Singh, Richa Garva, Marija Krstic-Demonacos, and Constantinos Demonacos. Cytochrome p450 2e1 (cyp2e1) regulates the response to oxidative stress and migration of breast cancer cells. *Breast Cancer Research*, 15:1–12, 2013.

Jing Li, Yanbo Chen, Hongyuan Yu, Jingshen Tian, Fengshun Yuan, Jialong Fan, Yupeng Liu, Lin Zhu, Fan Wang, Yashuang Zhao, et al. Dusp1 promoter methylation in peripheral blood leukocyte is associated with triple-negative breast cancer risk. *Scientific reports*, 7 (1):43011, 2017.

Quanfu Li, Yunkai Chu, Shengze Li, Liping Yu, Huayun Deng, Chunhua Liao, Xiaodong Liao, Chihyu Yang, Min Qi, Jinke Cheng, et al. The oncoprotein muc1 facilitates breast cancer progression by promoting pink1-dependent mitophagy via atad3a destabilization. *Cell Death & Disease*, 13(10):899, 2022.

Yangyang Li, Yan Guo, Fengzhi Chen, Yuqing Cui, Xuesong Chen, and Guangyue Shi. Male breast cancer differs from female breast cancer in molecular features that affect prognoses and drug responses. *Translational Oncology*, 45:101980, 2024.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

Theresa Link, Friederike Kuithan, Armin Ehninger, Jan Dominik Kuhlmann, Michael Kramer, Andreas Werner, Axel Gatzweiler, Barbara Richter, Gerhard Ehninger, Gustavo Baretton, et al. Exploratory investigation of psca-protein expression in primary breast cancer patients reveals a link to her2/neu overexpression. *Oncotarget*, 8(33):54592, 2017.

Wan Liu, Wenjing Wang, Ning Zhang, and Wen Di. The role of ccl20-ccr6 axis in ovarian cancer metastasis. *OncoTargets and therapy*, pages 12739–12750, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30.

Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Qi Luo, Jinsui Li, Xiaohan Su, Qiao Tan, Fangfang Zhou, and Shaoli Xie. Col11a1 serves as a biomarker for poor prognosis and correlates with immune infiltration in breast cancer. *Frontiers in Genetics*, 13:935860, 2022.

Xuemei Lv, Miao He, Yanyun Zhao, Liwen Zhang, Wenjing Zhu, Longyang Jiang, Yuanyuan Yan, Yue Fan, Hongliang Zhao, Shuqi Zhou, et al. Identification of potential key genes and pathways predicting pathogenesis and prognosis for triple-negative breast cancer. *Cancer cell international*, 19:1–12, 2019.

Andriana Margariti, Bernhard Winkler, Eirini Karamariti, Anna Zampetaki, Tsung neng Tsai, Dilair F. Baban, Jiannis Ragoussis, Yi Huang, Jing-Dong Jackie Han, Lingfang Zeng, Yanhua Hu, and Qingbo Xu. Direct reprogramming of fibroblasts into endothelial cells capable of angiogenesis and reendothelialization in tissue-engineered vessels. *Proceedings of the National Academy of Sciences*, 109:13793 – 13798, 2012. URL https://api.semanticscholar.org/CorpusID:21888753.

Louis TP Martin, Mark W Nachtigal, Tamara Selman, Elaine Nguyen, Jayme Salsman, Graham Dellaire, and Denis J Dupré. Bitter taste receptors are expressed in human epithelial ovarian and prostate cancers cells and noscapine stimulation impacts cell survival. *Molecular and Cellular Biochemistry*, 454(1):203–214, 2019.

Ti Meng, Leichao Liu, Ruifang Hao, Siying Chen, and Yalin Dong. Transgelin-2: a potential oncogenic factor. *Tumor Biology*, 39(6):1010428317702650, 2017.

Laura R Moffitt, Maree Bilandzic, Amy L Wilson, Yiqian Chen, Mark D Gorrell, Martin K Oehler, Magdalena Plebanski, and Andrew N Stephens. Hypoxia regulates dpp4 expression, proteolytic inactivation, and shedding from ovarian cancer cells. *International Journal of Molecular Sciences*, 21(21):8110, 2020.

Clément Morgat, Véronique Brouste, Adrien Chastel, Valérie Vélasco, Gaétan Macgrogan, and Elif Hindié. Expression of neurotensin receptor-1 (nts 1) in primary breast tumors, cellular distribution, and association with clinical and biological factors. *Breast Cancer Research and Treatment*, 190:403–413, 2021.

Tsuyoshi Morita and Ken'ichiro Hayashi. Tumor progression is mediated by thymosin-β4 through a tgfβ/mrtf signaling axis. *Molecular Cancer Research*, 16(5):880–893, 2018.

Maria Muccioli and Fabian Benencia. Toll-like receptors in ovarian cancer as targets for immunotherapies. *Frontiers in immunology*, 5:341, 2014.

Juan P Muñoz, Pablo Pérez-Moreno, Yasmín Pérez, and Gloria M Calaf. The role of micrornas in breast cancer and the challenges of their clinical application. *Diagnostics*, 13(19):3072, 2023.

Barzin Y Nabet, Yu Qiu, Jacob E Shabason, Tony J Wu, Taewon Yoon, Brian C Kim, Joseph L Benci, Angela M DeMichele, Julia Tchou, Joseph Marcotrigiano, et al. Exosome

rna unshielding couples stromal activation to pattern recognition receptor signaling in cancer. *Cell*, 170(2):352–366, 2017.

Ali Naderi, Andrew E Teschendorff, Juergen Beigel, Massimiliano Cariati, Ian O Ellis, James D Brenton, and Carlos Caldas. Bex2 is overexpressed in a subset of primary breast cancers and mediates nerve growth factor/nuclear factor-κb inhibition of apoptosis in breast cancer cell lines. *Cancer research*, 67(14):6725–6736, 2007.

S David Nathanson, Michael Detmar, Timothy P Padera, Lucy R Yates, Danny R Welch, Thomas C Beadnell, Adam D Scheid, Emma D Wrenn, and Kevin Cheung. Mechanisms of breast cancer metastasis. *Clinical & experimental metastasis*, 39(1):117–137, 2022.

Petra Kralj Novak, Nada Lavrač, Dragan Gamberger, and Antonija Krstačić. Csm-sd: Methodology for contrast set mining through subgroup discovery. *Journal of Biomedical Informatics*, 42(1):113–122, 2009.

Masanori Oshi, Yoshihisa Tokumaru, Fernando A Angarita, Li Yan, Ryusei Matsuyama, Itaru Endo, and Kazuaki Takabe. Degree of early estrogen response predict survival after endocrine therapy in primary and metastatic er-positive breast cancer. *Cancers*, 12(12): 3557, 2020.

Masanori Oshi, Lan Le, Fernando A Angarita, Yoshihisa Tokumaru, Li Yan, Ryusei Matsuyama, Itaru Endo, and Kazuaki Takabe. Association of allograft rejection response score with biological cancer aggressiveness and with better survival in triple-negative breast cancer (tnbc)., 2021.

Masanori Oshi, Ankit Patel, Rongrong Wu, Lan Le, Yoshihisa Tokumaru, Akimitsu Yamada, Li Yan, Ryusei Matsuyama, Takashi Ishikawa, Itaru Endo, et al. Enhanced immune response outperform aggressive cancer biology and is associated with better survival in triple-negative breast cancer. *NPJ Breast Cancer*, 8(1):92, 2022.

Raghu Pandurangi, Thillai Sekar, and Ramasamy Paulmurugan. Restoration of the lost human beta defensin-1 protein in cancer as a strategy to improve the efficacy of chemotherapy. *Journal of Medicinal Chemistry*, 67(16):14200–14209, 2024. doi: 10.1021/acs.jmedchem.4c01040. URL https://doi.org/10.1021/acs.jmedchem.4c01040. PMID: 39137365.

Mengxin Qi, Xianglan Yi, Baohui Yue, Mingxiang Huang, Sheng Zhou, and Jing Xiong. S100a6 inhibits mdm2 to suppress breast cancer growth and enhance sensitivity to chemotherapy. *Breast Cancer Research*, 25(1):55, 2023.

Zhiwei Qiao, Ying Jiang, Ling Wang, Lei Wang, Jing Jiang, and Jingru Zhang. Mutations in kiaa1109, cacna1c, bsn, akap13, celsr2, and helz2 are associated with the prognosis in endometrial cancer. *Frontiers in genetics*, 10:909, 2019.

Emad A Rakha, Gary M Tse, and Cecily M Quinn. An update on the pathological classification of breast cancer. *Histopathology*, 82(1):5–16, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Jose Manuel Sanchez-Lopez, Edna Ayerim Mandujano-Tinoco, Alfredo Garcia-Venzor, Laura Fatima Lozada-Rodriguez, Cecilia Zampedri, Salvador Uribe-Carvajal, Jorge Melendez-Zajgla, Vilma Maldonado, and Floria Lizarraga. Integrative analysis of transcriptional profile reveals linc00052 as a suppressor of breast cancer cell migration. *Cancer Biomarkers*, 30(4):365–379, 2021.

Natalia Sauer, Igor Matkowski, Grażyna Bodalska, Marek Murawski, Piotr Dziegiel, and Jacek Calik. Prognostic role of prolactin-induced protein (pip) in breast cancer. *Cells*, 12(18):2252, 2023.

Saurabh Sharma, Lakshay Malhotra, Paromita Mukherjee, Navneet Kaur, Thammineni Krishanlata, Chittur V Srikanth, Vandana Mishra, Basu Dev Banerjee, Abdul Samath Ethayathulla, and Radhey Shyam Sharma. Putative interactions between transthyretin and endosulfan ii and its relevance in breast cancer. *International Journal of Biological Macromolecules*, 235:123670, 2023.

Vivian Yvonne Shin, Jiawei Chen, Isabella Wai-Yin Cheuk, Man-Ting Siu, Chi-Wang Ho, Xian Wang, Hongchuan Jin, and Ava Kwong. Long non-coding rna neat1 confers oncogenic role in triple-negative breast cancer through modulating chemoresistance and cancer stemness. *Cell death & disease*, 10(4):270, 2019.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

Jean Ching-Yi Tien, Yali Zhai, Rong Wu, Yuping Zhang, Yu Chang, Yunhui Cheng, Abigail J Todd, Christina E Wheeler, Shuqin Li, Rahul Mannan, et al. Defining cdk12 as a tumor suppressor and therapeutic target in mouse models of tubo-ovarian high-grade serous carcinoma. *Proceedings of the National Academy of Sciences*, 122(24):e2426909122, 2025.

Wei Lue Tong, Yaping N Tu, Mohammad D Samy, Wade J Sexton, and George Blanck. Identification of immunoglobulin v (d) j recombinations in solid tumor specimen exome files: Evidence for high level b-cell infiltrates in breast cancer. *Human Vaccines & Immunotherapeutics*, 13(3):501–506, 2017.

Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Kunwar Somesh Vikramdeo, Shashi Anand, Sarabjeet Kour Sudan, Paramahansa Pramanik, Seema Singh, Andrew K Godwin, Ajay Pratap Singh, and Santanu Dasgupta. Profiling mitochondrial dna mutations in tumors and circulating extracellular vesicles of triple-negative breast cancer patients for potential biomarker development. *FASEB BioAdvances*, 5(10):412, 2023.

Vincent Vinh-Hung, Sue A Joseph, Nadege Coutty, Bevan Hong Ly, Georges Vlastos, and Nam Phong Nguyen. Age and axillary lymph node ratio in postmenopausal women with t1-t2 node positive breast cancer. *The oncologist*, 15(10):1050–1062, 2010.

Dujuan Wang, Guohong Liu, Balu Wu, Li Chen, Lihua Zeng, and Yunbao Pan. Clinical significance of elevated s100a8 expression in breast cancer patients. *Frontiers in oncology*, 8:496, 2018.

Runzhi Wang, Ronghua Wang, Jinjun Tian, Jian Wang, Huaxiao Tang, Tao Wu, and Hui Wang. Btg2 as a tumor target for the treatment of luminal a breast cancer. *Experimental and Therapeutic Medicine*, 23(5):1–11, 2022.

Yanyan Wang, Ning Sheng, Ying Xie, Sihan Chen, Jun Lu, Zifeng Zhang, Qun Shan, Dongmei Wu, Guihong Zheng, Mengqiu Li, et al. Low expression of crisp3 predicts a favorable prognosis in patients with mammary carcinoma. *Journal of Cellular Physiology*, 234(8):13629–13638, 2019.

Hans Wildiers, Ben Van Calster, Lonneke V van de Poll-Franse, Wouter Hendrickx, Jo Røislien, Ann Smeets, Robert Paridaens, Karen Deraedt, Karin Leunen, Caroline Weltens, et al. Relationship between age and axillary lymph node involvement in women with breast cancer. *Journal of clinical oncology*, 27(18):2931–2937, 2009.

Nicholas T Woods, Rafael D Mesquita, Michael Sweet, Marcelo A Carvalho, Xueli Li, Yun Liu, Huey Nguyen, C Eric Thomas, Edwin S Iversen Jr, Sylvia Marsillac, et al. Charting the landscape of tandem brct domain–mediated protein interactions. *Science signaling*, 5(242):rs6–rs6, 2012.

Carissa Wu, Sonali Parbhoo, Marton Havasi, and Finale Doshi-Velez. Learning optimal summaries of clinical time-series with concept bottleneck models. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 648–672. PMLR, 05–06 Aug 2022. URL https://proceedings.mlr.press/v182/wu22a.html.

Tian Xie, Shan Pan, Hang Zheng, Zilv Luo, Kingsley M Tembo, Muhammad Jamal, Zhongyang Yu, Yao Yu, Jing Xia, Qian Yin, et al. Peg10 as an oncogene: expression regulatory mechanisms and role in tumor progression. *Cancer cell international*, 18:1–10, 2018.

Si-Guang Xu, Pei-Jun Yan, and Zhi-Ming Shao. Differential proteomic analysis of a highly metastatic variant of human breast cancer cells using two-dimensional differential gel electrophoresis. *Journal of cancer research and clinical oncology*, 136:1545–1556, 2010.

Jie Yi, Liwen Ren, Dandan Li, Jie Wu, Wan Li, Guanhua Du, and Jinhua Wang. Trefoil factor 1 (tff1) is a potential prognostic biomarker with functional significance in breast cancers. *Biomedicine & Pharmacotherapy*, 124:109827, 2020.

Safoora Yousefi, Fatemeh Amrollahi, Mohamed Amgad, Chengliang Dong, Joshua E Lewis, Congzheng Song, David A Gutman, Sameer H Halani, Jose Enrique Velazquez Vega, Daniel J Brat, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1):1–11, 2017.

Wei Yu, Hongyan Chai, Ying Li, Haixia Zhao, Xianfei Xie, Hao Zheng, Chenlong Wang, Xue Wang, Guifang Yang, Xiaojun Cai, et al. Increased expression of cyp4z1 promotes tumor angiogenesis and growth in human breast cancer. *Toxicology and applied pharmacology*, 264(1):73–83, 2012.

Reihaneh Zarrizi, Martin R Higgs, Karolin Voßgröne, Maria Rossing, Birgitte Bertelsen, Muthiah Bose, Arne Nedergaard Kousholt, Heike Rösner, Bent Ejlertsen, Grant S Stewart, et al. Germline rbbp8 variants associated with early-onset breast cancer compromise replication fork stability. *The Journal of clinical investigation*, 130(8):4069–4080, 2020.

Xinhai Zhang, Chenglong Wang, Shujun Xia, Fei Xiao, Jianping Peng, Yuxuan Gao, Fengbin Yu, Chuandong Wang, and Xiaodong Chen. The emerging role of snornas in human disease. *Genes & Diseases*, 10(5):2064–2081, 2023.

Lizhe Zhu, Nan Ma, Bin Wang, Lei Wang, Can Zhou, Yu Yan, Jianjun He, and Yu Ren. Significant prognostic values of aquaporin mrna expression in breast cancer. *Cancer management and research*, pages 1503–1515, 2019.

Juan Zou, Yaokun Chen, Zeqi Ji, Danyi Liu, Xin Chen, Mengjia Chen, Kexun Chen, Haojia Lin, Yexi Chen, and Zhiyang Li. Identification of c4bpa as biomarker associated with immune infiltration and prognosis in breast cancer. *Translational Cancer Research*, 13 (1):25, 2024.

Appendix A. Results on Ovarian Cancer

We extended our experiments to ovarian cancer using RNA expression and clinical data from TCGA. The dataset consisted of 280 samples, and the top 1000 genes were selected by ranking them based on the index of dispersion. We evaluated the performance of four models: CPH, XGBoost, DeepSurv, and ConSurv-XGB. The results are summarized below:

- CPH: 0.56 ± 0.07
- XGBoost: 0.59 ± 0.05
- DeepSurv: 0.61 ± 0.05
- ConSurv-XGB: 0.61 ± 0.07

We observe that ConSurv-XGB performs competitively with both DeepSurv and XGBoost. Similar to the results on breast cancer, using the top 100 features in ConSurv-XGB yields the highest performance. Furthermore, the top two extracted concepts include genes known to be associated with tumor suppression and therapeutic relevance in ovarian cancer.

- **Concept 1**: *DDO* Chen et al. (2022), *CDH12* Tien et al. (2025), *BSND* Qiao et al. (2019) - These genes are associated with suppressing tumor growth.
- **Concept 2**: *DPP* Moffitt et al. (2020), *TAS2R8* Martin et al. (2019), *HS6ST2* Chen et al. (2023), *CCL20* Liu et al. (2020), *TLR7* Muccioli and Benencia (2014) - These genes have been shown to be associated with ovarian cancer and represent potential therapeutic targets.

Appendix B. Experimental details

B.1. Hyperparameter tuning and model training

To ensure optimal parameters for model training, we conducted hyperparameter tuning for CPH, XGBoost, RuleKit, DeepSurv, MLP, and ConSurv using grid search. Ten runs were performed, with data splits generated using different random seeds (run IDs and their corresponding seeds are listed in Appendix Table 3). For each run, the data was split into 70% training, 15% validation, and 15% test sets. Hyperparameter tuning was performed on the training and validation sets across all ten runs. The test sets were used to evaluate performance of the tuned models. The categorical features were one hot encoded and continuous features were min-max normalized.

The parameter combinations evaluated for each model are detailed in Appendix Table 4, with the best-performing values highlighted in the last column. For ConSurv, the optimal parameters from MLP were adopted, with only the learning rate tuned separately. For XGBoost, DeepSurv, MLP, and ConSurv models, early stopping was applied with a patience of 50 rounds (500 maximum iterations for XGBoost and 250 maximum epochs for DeepSurv, MLP and ConSurv). Adaptive Moment Estimation (Adam) Kingma and Ba (2015) was used for the gradient descent algorithm. The same training set used for training the ConSurv model was also used to train the corresponding XGBoost and RuleKit models for concept extraction.

Table 3: Run id with corresponding seed

Run id	Seed
R1	0
R2	7
R3	42
R4	200
R5	777
R6	999
R7	1303
R8	1995
R9	1996
R10	2405

B.2. Concordance Index

Concordance **I**ndex or CI (Uno et al. (2011); Harrell Jr et al. (1996)) is a used widely metric for evaluating performance of survival models. It is calculated as:

$$CI = \frac{N_c}{N_c + N_d} \tag{2}$$

Where, N_c is number of concordant pairs such that for patients i and j, $NN(x_i) < NN(x_j)$, $T_i > T_j$ and $e_j = 1$. N_d is number of discordant pairs such that for patients i and j, $NN(x_i) > NN(x_j)$, $T_i > T_j$ and $e_j = 1$. $NN(.)$ is predicted LogRisk, T is time-to-event and e is event variable.

B.3. Objective function: Cox-loss

Deep learning models for survival analysis such as AutoSurv and DeepSurv often employ Cox-loss for training the models. We used the same loss function in training ConSurv and MLP models. The loss function is as follows (from Jiang et al. (2024); Katzman et al. (2018)):

$$l(\theta) = -\frac{1}{N_{e=1}} \sum_{i:e_i=1} (NN(x_i) - \log \sum_{j \in R(T_i)} e^{NN(x_j)}) + \lambda ||\theta||_2^2 \tag{3}$$

where θ are model parameters, $NN(x_i)$ (LogRisk) is model output for input x_i, λ is l_2 regularization parameter, e is the event variable, $N_{e=1}$ are number of patients with an observable event and $R(T_i)$ are patients still at risk of failure at time t.

B.4. Set similarity indices

Jaccard similarity index for set A and B is given by Appendix Equation (4). It ranges from 0–1 with 0 indicating no overlap or similarity and 1 indicating perfect overlap or similarity.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

Table 4: Parameter list for hyperparameter tuning

Parameter	Values	Best value
	CPH	
penalizer	$[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$	0.3
l1_ratio	$[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$	0.1
	XGBoost	
max_depth	$[5, 10, 20, 40]$	10
learning_rate	$[10^{-1}, 10^{-2}, 10^{-3}]$	10^{-2}
n_estimator	$[10, 50, 100, 250, 500, 750, 1000, 2000]$	500
lambda	$[10^{-1}, 10^{-2}, 10^{-3}]$	10^{-2}
alpha	$[10^{-1}, 10^{-2}]$	10^{-1}
aft_loss_distribution	$[0.1, 0.5, 1.2]$	1.2
	RuleKit	
measures	$[c2, rss, correlation]$	*rss*
minsupp_new	$[3, 5, 7, 9, 11, 13]$	3
	DeepSurv	
lr	$[10^{-1}, 10^{-3}, 10^{-5}]$	10^{-5}
hidden_size	$[64, 128, 256]$	128
l2_reg	$[0.1, 0.01, 0.001]$	0.01
batch_size	$[32, 64]$	64
dropout	$[0, 0.1, 0.2]$	0.2
	MLP	
batch_size	$[16, 32, 64]$	64
L2_reg	$[10^{-1}, 10^{-2}, 10^{-3}]$	10^{-3}
learning_rate	$[10^{-1}, 10^{-3}, 10^{-5}]$	10^{-5}
hidden_size	$[32, 64, 128, 256, 512]$	64
n_layer	$[1, 2, 3]$	2
	ConSurv	
learning_rate	$[10^{-1}, 10^{-3}, 10^{-5}]$	10^{-3}

Cosine similarity index for set A and B is given by Appendix Equation (5). First the features from the two concepts being compared are one hot encoded into vectors a and b and then the Cosine similarity is calculated. The range of Cosine similarity is from -1–1, however, in our case of binary vectors, it ranges only from 0–1.

$$Cosine(A, B) = \frac{|a \cdot b|}{||a||||b||} \tag{5}$$

Appendix C. Contrast Set Mining

Association rule mining is used in problems with multi-dimensional datasets to identify relations between features (Agrawal et al. (1993)). These associations takes a form of `IF-THEN` rules. Contrast set mining (Bay and Pazzani (2001); Novak et al. (2009); Gudyś et al. (2024)) is a special type of association rule mining that identifies key differences between groups within a dataset. It specifically aims to find attribute-value (i.e. feature and its value) combinations, known as contrast sets, that vary significantly in frequency or magnitude across these groups.

For a dataset $\mathcal{D}$ containing two classes—positive (P) and negative (N)—a contrast set S is defined as a combination of attribute-value pairs (e.g., $A_1 = v_1 \wedge A_2 = v_2$, where v_1 and v_2 are specific values of the attributes). The support of S within positive or negative class, represented as $\text{support}(S, P)$ or $\text{support}(S, N)$, is the proportion of instances in class that satisfy S. A contrast set S is deemed significant if there is at least one pair of classes, P and N, for which the support difference $|\text{support}(S, P) - \text{support}(S, N)|$ exceeds a given threshold δ and is statistically significant. For multiclass problems, it employs a one-vs-all strategy to generate rules specific to each class. In this work, we leverage the `RuleKit` (Gudyś et al. (2020, 2024)) package to extract contrast sets.

Survival classification using Rule-based model: As an additional means for pattern extraction used in our study, we trained RuleKit, a contrast set rule-based model. We transformed our problem into a classification task by categorizing patients into three groups based on survival times as standard practice Grootes et al. (2024): < 5 years, 5–10 years, and > 10 years. The classification AUROC were 0.59 (< 5 years), 0.6 (5–10 years) and 0.55 (> 10 years). Each run produced between 61–83 rules, with median rule length (or number of features per rule) of 21. About 75% of rules from each run belonged to class < 5 survival years, 20% belonged to class 5–10 survival years and remaining 5% belonged to class > 10 survival years. Across runs, 44 features were consistently used by RuleKit (See Appendix Table 5). Since the primary purpose of splitting patients into three groups was to identify and group features (concepts) associated with long-term or short-term survival, rather than to predict risk, censoring data is treated as a lower bound when assigning survival classes.

Appendix D. Biological interpretation of concepts

D.1. Gene Set Variation Analysis

Gene **S**et **V**ariation **A**nalysis (GSVA) is a non-parametric, unsupervised method that estimates the variation of pathway activity (or gene set enrichment) across all samples in a transcriptomic dataset (Hänzelmann et al. (2013)). Rather than examining individual gene expression, GSVA computes an enrichment score per sample for each predefined gene set, effectively transforming the gene-expression matrix into a sample-by-gene-set score matrix. In our work, we used GSVA to aggregate the genes associated with each concept (rule) into a single "concept activity score" per patient. The aggregated GSVA scores enable stratification of patients into survival groups, which we visualize using Kaplan-Meier plots.

D.2. Over-Representation Analysis

We performed **O**ver-**R**epresentation **A**nalysis (ORA) to identify biological pathways or gene sets enriched among top genes from each ConSurv rule. Using the **M**olecular **Sig**natures **D**ata**B**ase (MSigDB) collections (Liberzon et al. (2015)) (H - Hallmark gene sets and C6 - oncogenic signatures) and all measured genes as a background set, we applied Fisher's exact test to evaluate whether a given genes in a concept appeared more frequently in any particular gene set than expected by chance. We corrected for multiple testing using the Benjamini–Hochberg method (FDR < 0.05). These enriched pathways provide functional context for each rule, underscoring potential mechanisms underlying risk stratification.

D.3. Biological relevance of the common features:

To determine whether the consistently relevant common genes of XGBoost and RuleKit play a central role in survival outcomes, we performed Gene Set Variation Analysis (GSVA, Hänzelmann et al. (2013)). GSVA enables a non-parametric assessment of gene set enrichment. The dataset specifications for running GSVA are provided in Appendix D.1. Appendix Figure 6 illustrate the survival stratification based on common genes for both models. For XGBoost and RuleKit, a significant stratification between high-risk and low-risk groups is observed ($p = 0.00093$ and $p = 0.0029$, respectively). GSVA further reveals that the common genes identified by XGBoost are associated with the Gene Ontology (GO) molecular function GO:0005198 (i.e., structural molecule activity), which is relatively non-specific. In contrast, the common genes identified by RuleKit are linked to a more specific biological process—programmed cell death (GO biological process GO:0012501). The ability of both models to leverage common genes for survival stratification underscores their effectiveness in identifying core biological features that contribute to predictive performance.

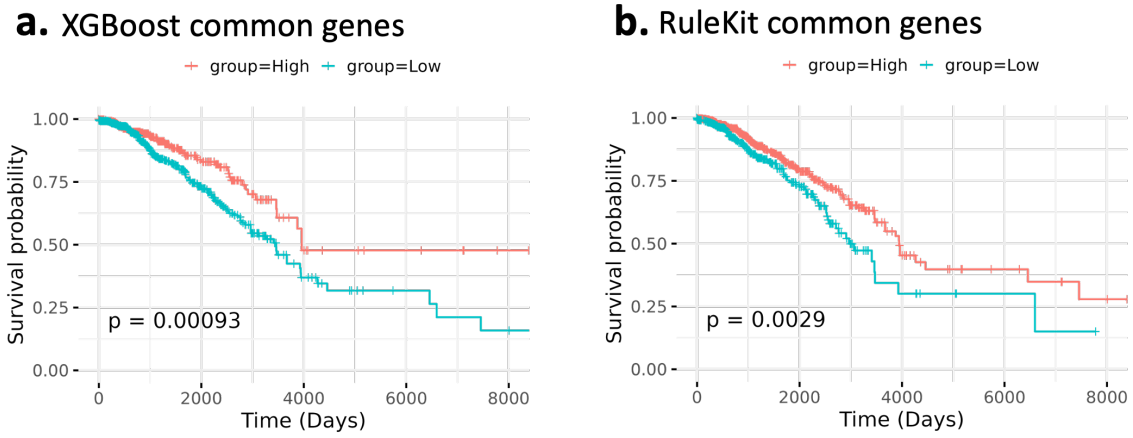

Figure 6: **KM plots for common genes selected by XGBoost and RuleKit** shows significant stratification between high- and low-risk groups ($p = 0.00093$ and $p = 0.0029$ respectively).

Table 5: Common features across 10 different runs from XGBoost and RuleKit (Genes converted using https://www.biotools.fr/)

Feature	Other name	Reference
	XGBoost	
ENSG00000183666.17	GUSBP1	Woods et al. (2012)
ENSG00000272398.6	CD24	Chan et al. (2024)
ENSG00000164879.7	CA3	Lee et al. (2023)
ENSG00000143631.11	FLG	Adnane et al. (1991)
ENSG00000105388.16	CEACAM5	Bechmann et al. (2020)
ENSG00000185275.6	CD24P4	Kristiansen et al. (2003)
ENSG00000109072.14	VTN	Bera et al. (2020)
ENSG00000236824.2	BCYRN1	Booy et al. (2017)
ENSG00000170835.16	CEL	Cui et al. (2019)
ENSG00000120129.6	DUSP1	Li et al. (2017); Candas et al. (2014)
ENSG00000129824.16	RPS4Y1	Li et al. (2024)
days_to_birth	Age	Fernandes et al. (2023); Brandt et al. (2015)
ajcc_pathologic_m_M1	Metastasis pathology	Rakha et al. (2023)
ajcc_pathologic_n_N1b	Lymph nodes pathology	Rakha et al. (2023)
	RuleKit	
ENSG00000276168.1	RN7SL1	Nabet et al. (2017)
ENSG00000278771.1	RN7SL3	-
ENSG00000171201.12	SMR3B	Lv et al. (2019)
ENSG00000159763.4	PIP	Sauer et al. (2023)
ENSG00000133636.11	NTS	Morgat et al. (2021)
ENSG00000089199.10	CHGB	Kimura et al. (2002)
ENSG00000143556.9	S100A7	Emberley et al. (2004)
ENSG00000143546.10	S100A8	Wang et al. (2018)
ENSG00000118271.12	TTR	Sharma et al. (2023)
ENSG00000172551.11	MUCL1	Li et al. (2022)
ENSG00000198888.2	MT-ND1	Vikramdeo et al. (2023)
ENSG00000212283.1	SNORD89	Zhang et al. (2023)
ENSG00000221716.1	SNORA11	-
ENSG00000145824.13	CXCL14	Gu et al. (2012)
ENSG00000126709.15	IFI6	Cheriyath et al. (2018)
ENSG00000225972.1	MTND1P23	-
ENSG00000164879.7	CA3	Lee et al. (2023)
ENSG00000211637.2	IGLV4-69	Tong et al. (2017)
ENSG00000210196.2	MT-TP	-
ENSG00000130649.10	CYP2E1	Leung et al. (2013)
ENSG00000227234.2	SPANXB1	Kannan et al. (2019)
ENSG00000133169.6	BEX1	Naderi et al. (2007)
ENSG00000273716.2	AC092670.1	-
ENSG00000187653.11	TMSB4XP8	Morita and Hayashi (2018)
ENSG00000159388.6	BTG2	Wang et al. (2022)
ENSG00000176907.5	TCIM	-
ENSG00000224543.4	SNRPGP15	-
ENSG00000276699.1	TRAJ36	-
ENSG00000242265.6	PEG10	Xie et al. (2018)
ENSG00000133063.16	CHIT1	-
ENSG00000259527.2	LINC00052	Sanchez-Lopez et al. (2021)
ENSG00000229344.1	MTCO2P12	-
ENSG00000101470.10	TNNC2	-
ENSG00000167653.5	PSCA	Link et al. (2017)
ENSG00000105894.12	PTN	Ganguly et al. (2023)
ENSG00000096006.12	CRISP3	Wang et al. (2019)
ENSG00000123838.11	C4BPA	Zou et al. (2024)
ENSG00000198763.3	MT-ND2	-
ENSG00000251562.8	MALAT1	Kim et al. (2018); Arun and Spector (2019)
ENSG00000186160.5	CYP4Z1	Yu et al. (2012)
ENSG00000161634.12	DCD	Brauer et al. (2014)
ENSG00000175426.11	PCSK1	-
days_to_birth	Age	Fernandes et al. (2023); Brandt et al. (2015)
ajcc_pathologic_stage I	Pathology stage	Rakha et al. (2023)

Appendix E. Quantitative analysis of the interpretable concepts

In this section, we analyze the quantitative aspects of the interpretable concepts in our framework, starting with an exploratory analysis (Appendix Figure 7) and then examining the stability and robustness of concept extraction. Appendix Figure 7a shows that XGBoost uses 35–50% of the 1,056 features per run, producing 330–390 trees (Appendix Figure 7b). The tree depth affects the number of unique concepts, with a significant drop from depth 10 to depth 1 (Appendix Figure 7c). Since pruning at depth 5 yields comparable performance to depth 10, we extract concepts at this depth, resulting in 540–770 unique concepts (Appendix Figure 7d).

In RuleKit, 43–53% of features are utilized, producing 61–83 rules, each containing 25–30 conditions (Appendix Figure 7e–g). Unlike XGBoost, each rule corresponds to a unique concept, resulting in 61–83 unique concepts per run (Appendix Figure 7h).

The top 100 concepts, which are key for ConSurv-XGB's predictive capability, were analyzed for their origin. Figure 8a shows most high-ranking concepts emerge in the initial and final training iterations. Figure 8b reveals that larger concepts, particularly those with five features, dominate the top 100. Figure 8d demonstrates that ConSurv-Rule has more stable concept ranking than Consurv-XGB.

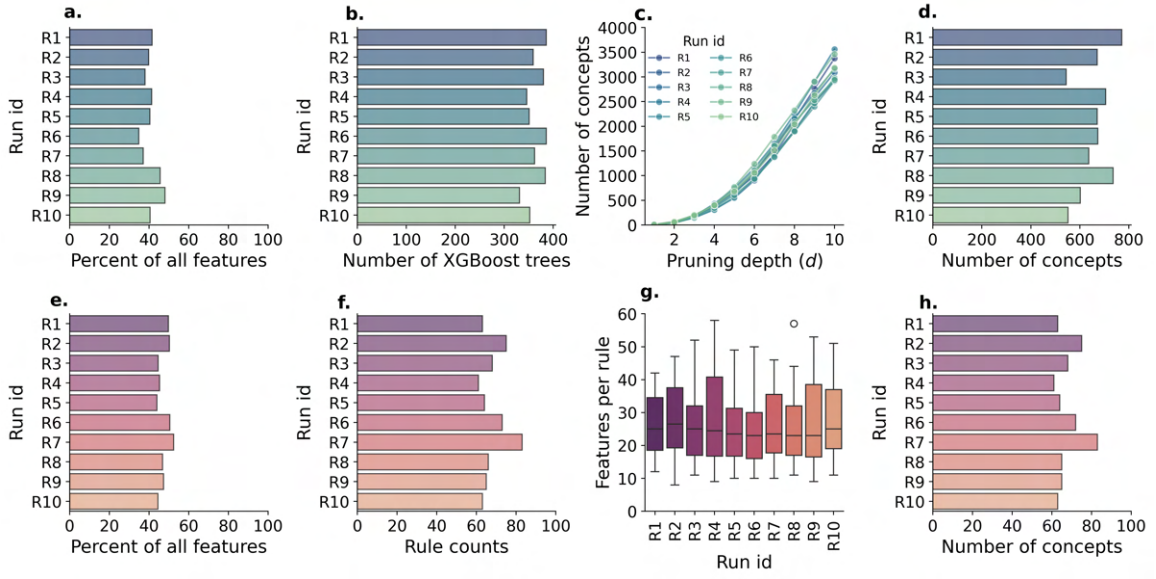

Figure 7: **XGBoost uses fewer features but generates more concepts than RuleKit.** XGBoost generates 330–390 trees using 35–50% of features per run. Pruning the XGBoost trees drops concept counts from 3,000 (depth 10) to fewer than 10 (depth 1). RuleKit generates 61–83 rules (with 25–30 conditions per rule), covering 43–53% of features per run. Each rule roughly translates to one unique concept.

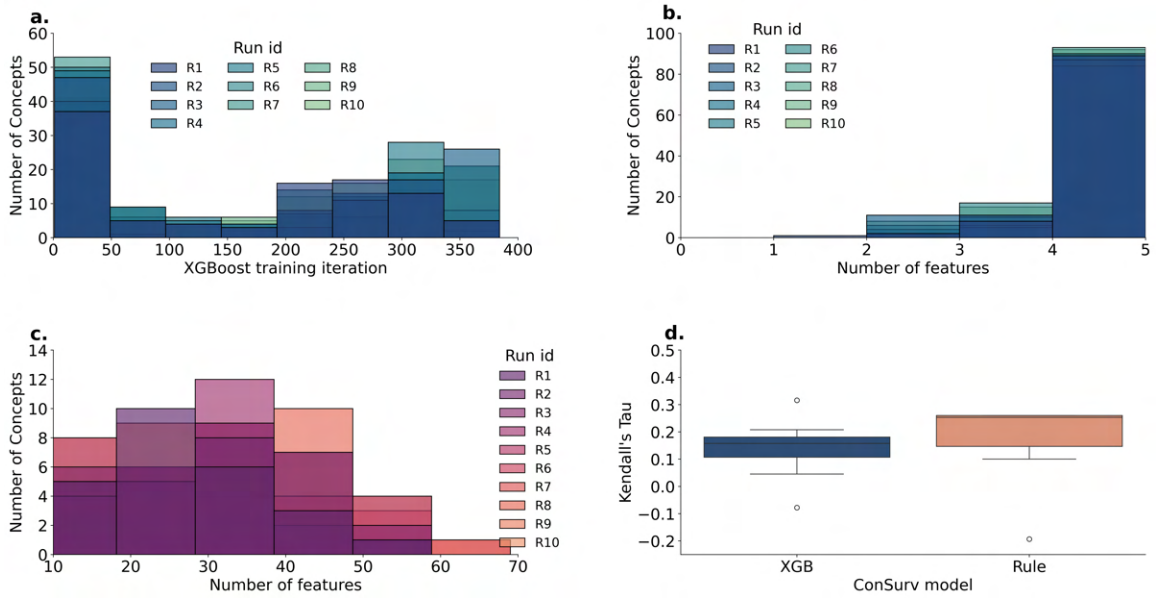

Figure 8: Most top 100 ConSurv-XGB concepts originate from early or late training iterations (a). Most ConSurv-XGB tend to contain 5 features, whereas ConSurv-Rule tend to contain 20–40 features (b,c). Concepts in ConSurv-Rule appear to be more stable than ConSurv-XGB, but overall the rankings of the top concepts (100 for -XGB and 25 for -Rule) change after retraining across runs (d).

Table 6: Mann–Whitney U test results on test CI distribution

model	**model**	**statistic**	**p-value**
C-Rule Top25	C-XGB Top100	46.0	0.791337
C-Rule Top25	CPH	78.0	**0.037635**
C-Rule Top25	DeepSurv	36.0	0.307489
C-Rule Top25	MLP	37.0	0.344704
C-Rule Top25	XGBoost	47.0	0.850107
C-XGB Top100	CPH	84.0	**0.011330**
C-XGB Top100	DeepSurv	42.0	0.570750
C-XGB Top100	MLP	42.0	0.570750
C-XGB Top100	XGBoost	55.0	0.733730
CPH	DeepSurv	17.0	**0.014019**
CPH	MLP	16.0	**0.011330**
CPH	XGBoost	31.0	0.161972
DeepSurv	MLP	54.0	0.791337
DeepSurv	XGBoost	60.0	0.472676
MLP	XGBoost	58.0	0.570750

Table 7: ConSurv-XGB top 5 concepts for Run 6

Features	Alternative Name
	Concept 1, Rank 1
demographic.days_to_birth	Age
ENSG00000283498.1	MIR1244-2
ENSG00000211660.3	IGLV2-23
ajcc_pathologic_n_N0 (i-)	Lymph node pathology
	Concept 2, Rank 2
ENSG00000245532.9	NEAT1
ENSG00000160182.3	TFF1
ENSG00000230667.6	SETSIP
ENSG00000197956.10	S100A6
ENSG00000129824.16	RPS4Y1
	Concept 3, Rank 3
ENSG00000103375.11	AQP8
ENSG00000175426.11	PCSK1
ENSG00000106211.10	HSPB1
ENSG00000171747.9	LGALS4
ajcc_pathologic_n_N2	Lymph node pathology
	Concept 4, Rank 4
ENSG00000175445.17	LPL
ENSG00000204248.11	COL11A2
ENSG00000153002.12	CPB1
ENSG00000158710.15	TAGLN2
ajcc_pathologic_n_N1b	Lymph node pathology
	Concept 5, Rank 5
ENSG00000211966.2	IGHV5-51
ENSG00000129824.16	RPS4Y1
ENSG00000223609.11	HBD
ENSG00000185275.6	CD24P4
days_to_birth	Age

Table 8: ConSurv-Rule Concept 1, Rank 1 for Run 6

Features	Alternative Name
ENSG00000197249.14	SERPINA1
ENSG00000146648.19	EGFR
ENSG00000141424.13	SLC39A6
ENSG00000170345.10	FOS
ENSG00000225972.1	MTND1P23
ENSG00000189058.9	APOD
ENSG00000172551.11	MUCL1
ENSG00000134827.8	TCN1
ENSG00000211639.2	IGLV4-60
ENSG00000141750.7	STAC2
ENSG00000134258.17	VTCN1
ENSG00000122711.9	SPINK4
ENSG00000170421.12	KRT8
ENSG00000160182.3	TFF1
ENSG00000147256.12	ARHGAP36
ENSG00000054938.16	CHRDL2
ENSG00000211637.2	IGLV4-69
ENSG00000204434.6	POTEKP
ENSG00000263934.5	SNORD3A
ENSG00000253818.1	IGLV1-41
ENSG00000138696.11	BMPR1B
ENSG00000263639.7	MSMB
ENSG00000167754.13	KLK5
ENSG00000106541.12	AGR2
ENSG00000159335.18	PTMS
ENSG00000198938.2	MT-CO3
ENSG00000197253.13	TPSB2
ENSG00000139329.5	LUM
ENSG00000134201.12	GSTM5
ENSG00000163736.4	PPBP
ENSG00000211660.3	IGLV2-23
ENSG00000012660.14	ELOVL5
ENSG00000123999.5	INHA
ENSG00000253802.2	AC105999.2
ENSG00000146678.10	IGFBP1
ENSG00000164120.14	HPGD
ENSG00000113739.10	STC2
ENSG00000094755.17	GABRP
ENSG00000200087.1	SNORA73B
ENSG00000112306.8	RPS12
ENSG00000167258.15	CDK12
ENSG00000178795.9	GDPD4
ENSG00000286775.1	-
ENSG00000164756.12	SLC30A8
ENSG00000171747.9	LGALS4
ENSG00000222880.1	RN7SKP261
ENSG00000109321.11	AREG
ENSG00000211655.3	IGLV1-36
ENSG00000163631.17	ALB
ENSG00000141232.5	TOB1
ENSG00000143125.6	PROK1
ENSG00000108679.13	LGALS3BP
ENSG00000172724.12	CCL19
ENSG00000173335.5	CST9
ENSG00000187653.11	TMSB4XP8
ENSG00000179593.16	ALOX15B
ENSG00000211950.2	IGHV1-24
ENSG00000140988.16	RPS2
ajcc_pathologic_stage I	Pathology stage
ajcc_pathologic_stage IA	Pathology stage
ajcc_pathologic_stage IB	Pathology stage
ajcc_pathologic_stage II	Pathology stage
ajcc_pathologic_stage IIA	Pathology stage
ajcc_pathologic_stage IIB	Pathology stage
ajcc_pathologic_stage III	Pathology stage
ajcc_pathologic_stage IIIA	Pathology stage
ajcc_pathologic_stage IIIB	Pathology stage
ajcc_pathologic_stage IIIC	Pathology stage
ajcc_pathologic_stage IV	Pathology stage
ajcc_pathologic_stage X	Pathology stage

Table 9: ConSurv-Rule Concept 2, Rank 2 for Run 6

Features	Alternative Name
ENSG00000167768.4	KRT1
ENSG00000128016.7	ZFP36
ENSG00000105894.12	PTN
ENSG00000170893.4	TRH
ENSG00000200164.1	RF00019
ENSG00000140986.8	RPL3L
ENSG00000218175.2	AC016739.1
ENSG00000105388.16	CEACAM5
ENSG00000199629.1	RNU1-14P
ENSG00000173821.19	RNF213
ENSG00000238034.1	AL109807.1
ENSG00000241351.3	IGKV3-11
ENSG00000165949.12	IFI27
ENSG00000137673.9	MMP7
ENSG00000109072.14	VTN
ENSG00000147604.14	RPL7
ENSG00000161634.12	DCD
ENSG00000099194.6	SCD
ENSG00000286339.1	-
ENSG00000207205.1	RNVU1-15
ENSG00000115414.21	FN1
ENSG00000212283.1	SNORD89
ENSG00000171246.6	NPTX1
ENSG00000198888.2	MT-ND1
ENSG00000211866.1	TRAJ23
ENSG00000204936.10	CD177
ENSG00000229344.1	MTCO2P12
ENSG00000104267.10	CA2
ENSG00000163220.11	S100A9
ENSG00000121769.8	FABP3
ENSG00000229859.10	PGA3
ENSG00000240409.1	MTATP8P1
ENSG00000133048.13	CHI3L1
ENSG00000211598.2	IGKV4-1
ENSG00000281990.1	IGHV1-69-2
ENSG00000206503.13	HLA-A
ENSG00000212605.1	RNU1-56P
ENSG00000273716.2	AC092670.1
ENSG00000166710.21	B2M
ENSG00000096006.12	CRISP3
ENSG00000075388.4	FGF4
ENSG00000242265.6	PEG10
ENSG00000161055.4	SCGB3A1
ENSG00000162267.12	ITIH3
ENSG00000240216.7	CPHL1P
ENSG00000151224.13	MAT1A

Table 10: ConSurv-Rule Concept 3, Rank 3 for Run 6

Features	Alternative Name
ENSG00000171428.15	NAT1
ENSG00000170345.10	FOS
ENSG00000178473.7	UCN3
ENSG00000136881.12	BAAT
ENSG00000182853.12	VMO1
ENSG00000189058.9	APOD
ENSG00000141736.14	ERBB2
ENSG00000170807.12	LMOD2
ENSG00000175899.15	A2M
ENSG00000183666.17	GUSBP1
ENSG00000129988.6	LBP
ENSG00000116882.15	HAO2
ENSG00000141750.7	STAC2
ENSG00000132703.4	APCS
ENSG00000278233.1	RNA5-8SN3
ENSG00000054938.16	CHRDL2
ENSG00000156885.6	COX6A2
ENSG00000124107.5	SLPI
ENSG00000133110.15	POSTN
ENSG00000212907.2	MT-ND4L
ENSG00000159388.6	BTG2
ENSG00000248746.6	ACTN3
ENSG00000265929.1	MIR5195
ENSG00000263639.7	MSMB
ENSG00000118849.10	RARRES1
ENSG00000228253.1	MT-ATP8
ENSG00000174697.5	LEP
ENSG00000211976.2	IGHV3-73
ENSG00000016490.16	CLCA1
ENSG00000197253.13	TPSB2
ENSG00000142089.16	IFITM3
ENSG00000248144.6	ADH1C
ENSG00000143248.13	RGS5
ENSG00000005381.8	MPO
ENSG00000249780.1	AC093809.1
ENSG00000012660.14	ELOVL5
ENSG00000248779.1	AC093297.1
ENSG00000141753.7	IGFBP4
ENSG00000253755.1	IGHGP
ENSG00000186009.5	ATP4B
ENSG00000183607.10	GKN2
ENSG00000109072.14	VTN
ENSG00000261409.1	AL035425.3
ENSG00000158104.11	HPD
ENSG00000124939.6	SCGB2A1
ENSG00000225630.1	MTND2P28
ENSG00000102837.7	OLFM4
ENSG00000131771.14	PPP1R1B
ENSG00000164128.7	NPY1R
ENSG00000101443.18	WFDC2
ENSG00000140459.18	CYP11A1
ENSG00000172724.12	CCL19
ENSG00000179593.16	ALOX15B
ENSG00000159167.12	STC1
days_to_birth	Age

Table 11: ConSurv-Rule Concept 4, Rank 4 for Run 6

Features	Alternative Name
ENSG00000125999.11	BPIFB1
ENSG00000166426.8	CRABP1
ENSG00000211668.2	IGLV2-11
ENSG00000104368.19	PLAT
ENSG00000092054.13	MYH7
ENSG00000164326.5	CARTPT
ENSG00000143632.14	ACTA1
ENSG00000019582.15	CD74
ENSG00000205361.8	MT1DP
ENSG00000229314.5	ORM1
ENSG00000239855.1	IGKV1-6
ENSG00000240386.3	LCE1F
ENSG00000248527.1	MTATP6P1
ENSG00000170367.5	CST5
ENSG00000283907.1	AD000090.1
ENSG00000230715.3	AC018638.2
ENSG00000096384.20	HSP90AB1
ENSG00000241755.1	IGKV1-9
ENSG00000171246.6	NPTX1
ENSG00000204936.10	CD177
ENSG00000171345.13	KRT19
ENSG00000145824.13	CXCL14
ENSG00000124107.5	SLPI
ENSG00000167676.4	PLIN4
ENSG00000169347.17	GP2
ENSG00000141744.4	PNMT
ENSG00000221421.1	MIR1283-1
ENSG00000231414.1	AC016700.2
ENSG00000118271.12	TTR
ENSG00000244461.1	LINC02077
ENSG00000132693.12	CRP
ENSG00000167531.6	LALBA
ajcc_pathologic_n_N0	Lymph node pathology
ajcc_pathologic_n_N0 (i+)	Lymph node pathology
ajcc_pathologic_n_N0 (i-)	Lymph node pathology
ajcc_pathologic_n_N0 (mol+)	Lymph node pathology
ajcc_pathologic_n_N1	Lymph node pathology
ajcc_pathologic_n_N1a	Lymph node pathology
ajcc_pathologic_n_N1b	Lymph node pathology
ajcc_pathologic_n_N1c	Lymph node pathology
ajcc_pathologic_n_N1mi	Lymph node pathology
ajcc_pathologic_n_N2	Lymph node pathology
ajcc_pathologic_n_N2a	Lymph node pathology
ajcc_pathologic_n_N3	Lymph node pathology
ajcc_pathologic_n_N3a	Lymph node pathology
ajcc_pathologic_n_N3b	Lymph node pathology
ajcc_pathologic_n_N3c	Lymph node pathology
ajcc_pathologic_n_NX	Lymph node pathology

Table 12: ConSurv-Rule Concept 5, Rank 5 for Run 6

Features	**Alternative Name**
ENSG00000186847.6	KRT14
ENSG00000248144.6	ADH1C
ENSG00000263426.2	RN7SL471P
ENSG00000139329.5	LUM
ENSG00000211890.4	IGHA2
ENSG00000197249.14	SERPINA1
ENSG00000143248.13	RGS5
ENSG00000198899.2	MT-ATP6
ENSG00000171428.15	NAT1
ENSG00000170345.10	FOS
ENSG00000096088.16	PGC
ENSG00000171346.16	KRT15
ENSG00000136881.12	BAAT
ENSG00000136929.13	HEMGN
ENSG00000196296.14	ATP2A1
ENSG00000211662.2	IGLV3-21
ENSG00000164404.8	GDF9
ENSG00000129988.6	LBP
ENSG00000167757.14	KLK11
ENSG00000134827.8	TCN1
ENSG00000253755.1	IGHGP
ENSG00000141750.7	STAC2
ENSG00000135413.9	LACRT
ENSG00000243302.3	AC018638.4
ENSG00000109072.14	VTN
ENSG00000086967.10	MYBPC2
ENSG00000170421.12	KRT8
ENSG00000147256.12	ARHGAP36
ENSG00000131771.14	PPP1R1B
ENSG00000124102.5	PI3
ENSG00000145192.13	AHSG
ENSG00000265972.6	TXNIP
ENSG00000163220.11	S100A9
ENSG00000133048.13	CHI3L1
ENSG00000159388.6	BTG2
ENSG00000231500.7	RPS18
ENSG00000263639.7	MSMB
ENSG00000198786.2	MT-ND5
ENSG00000265929.1	MIR5195
ENSG00000211976.2	IGHV3-73
days_to_birth	Age

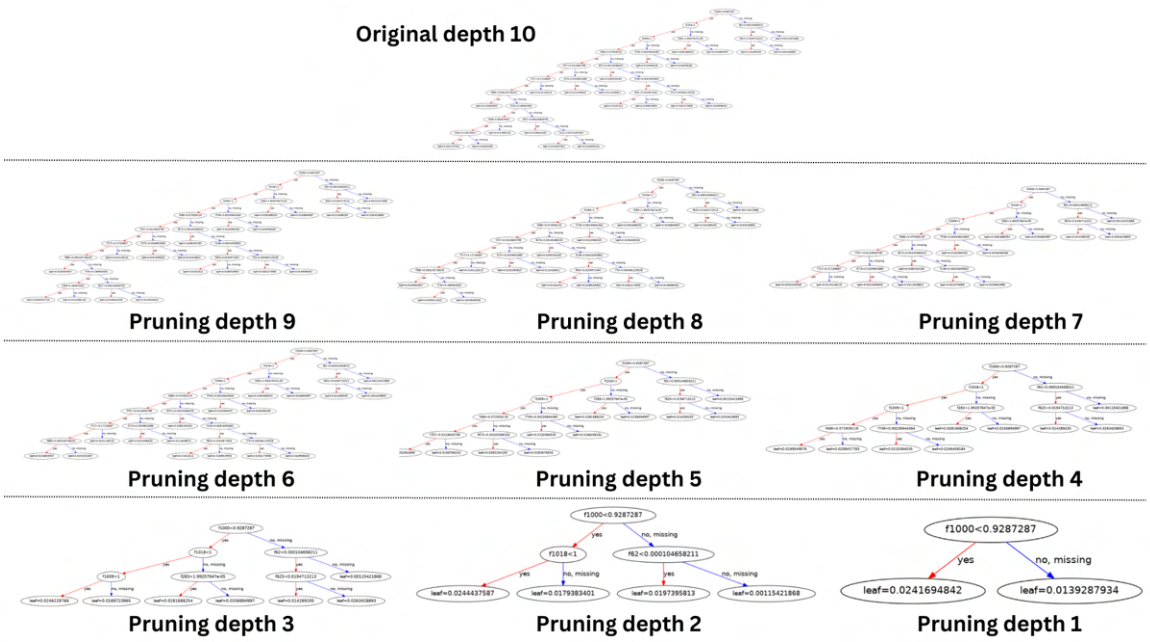

Figure 9: Pruning an XGBoost tree from original depth of 10 to a pruning depth of 1

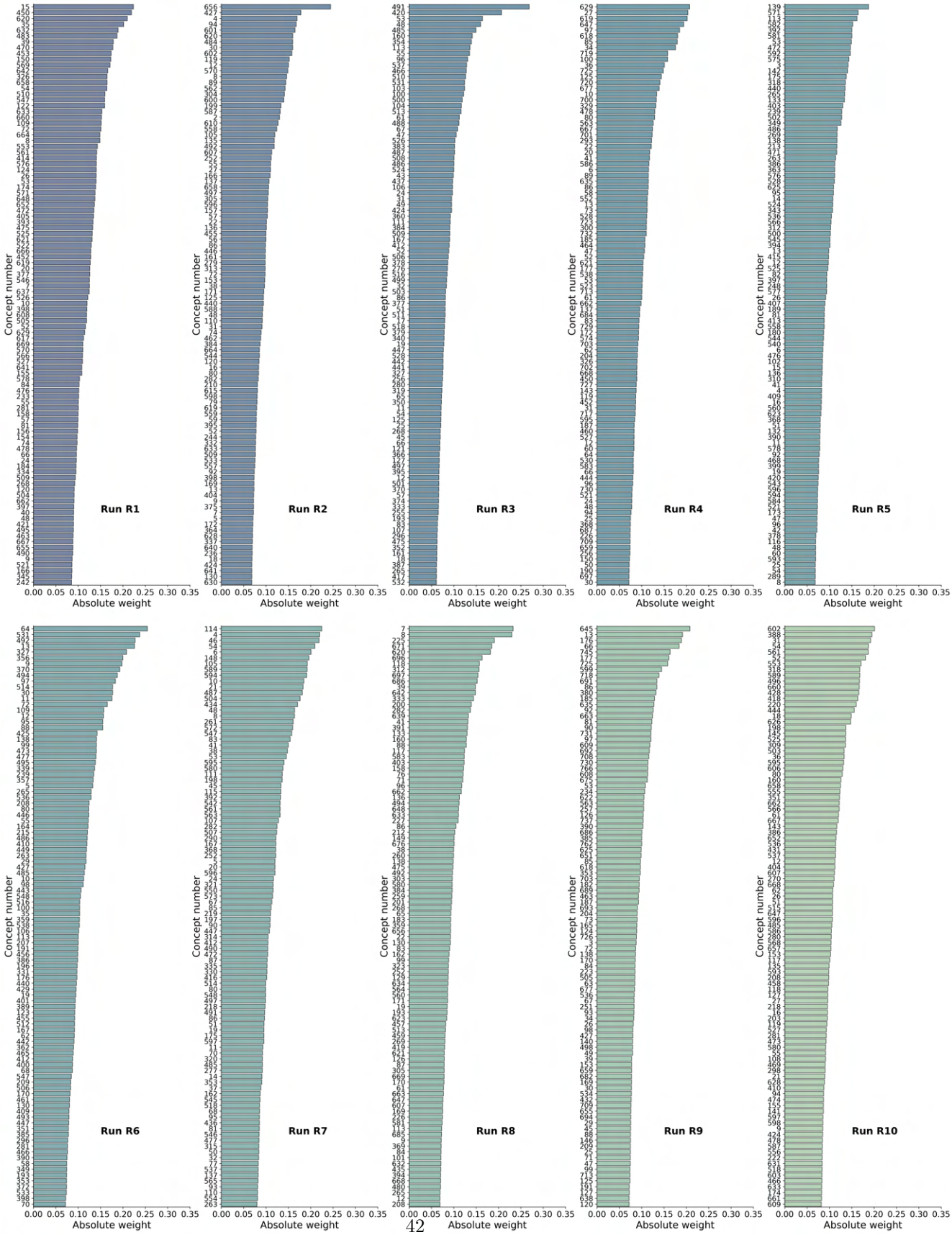

Figure 10: Top 100 concepts ordered by absolute weights in ConSurv-XGB-all across all 10 runs

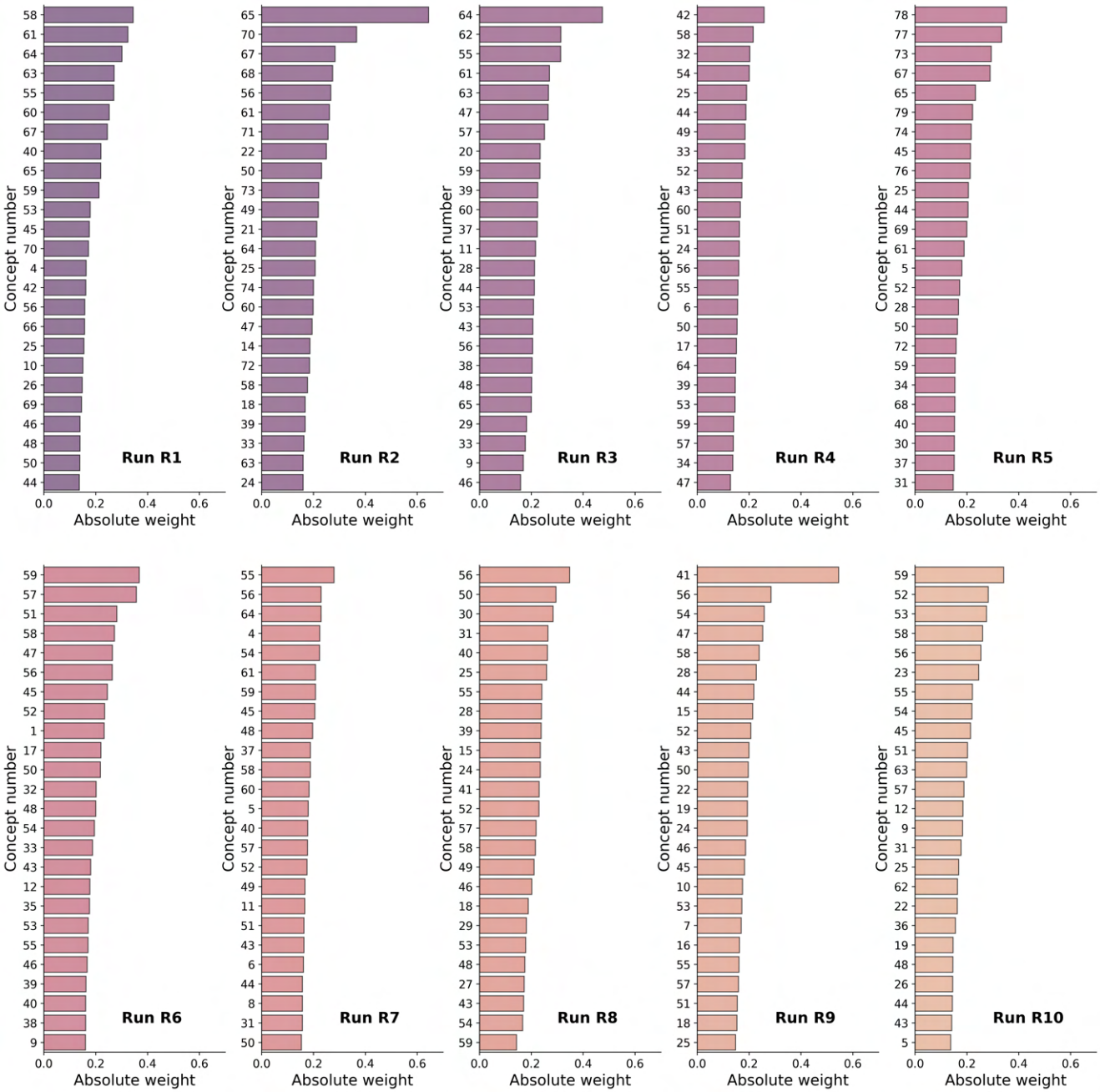

Figure 11: Top 25 concepts ordered by absolute weights in ConSurv-Rule-all across all 10 runs

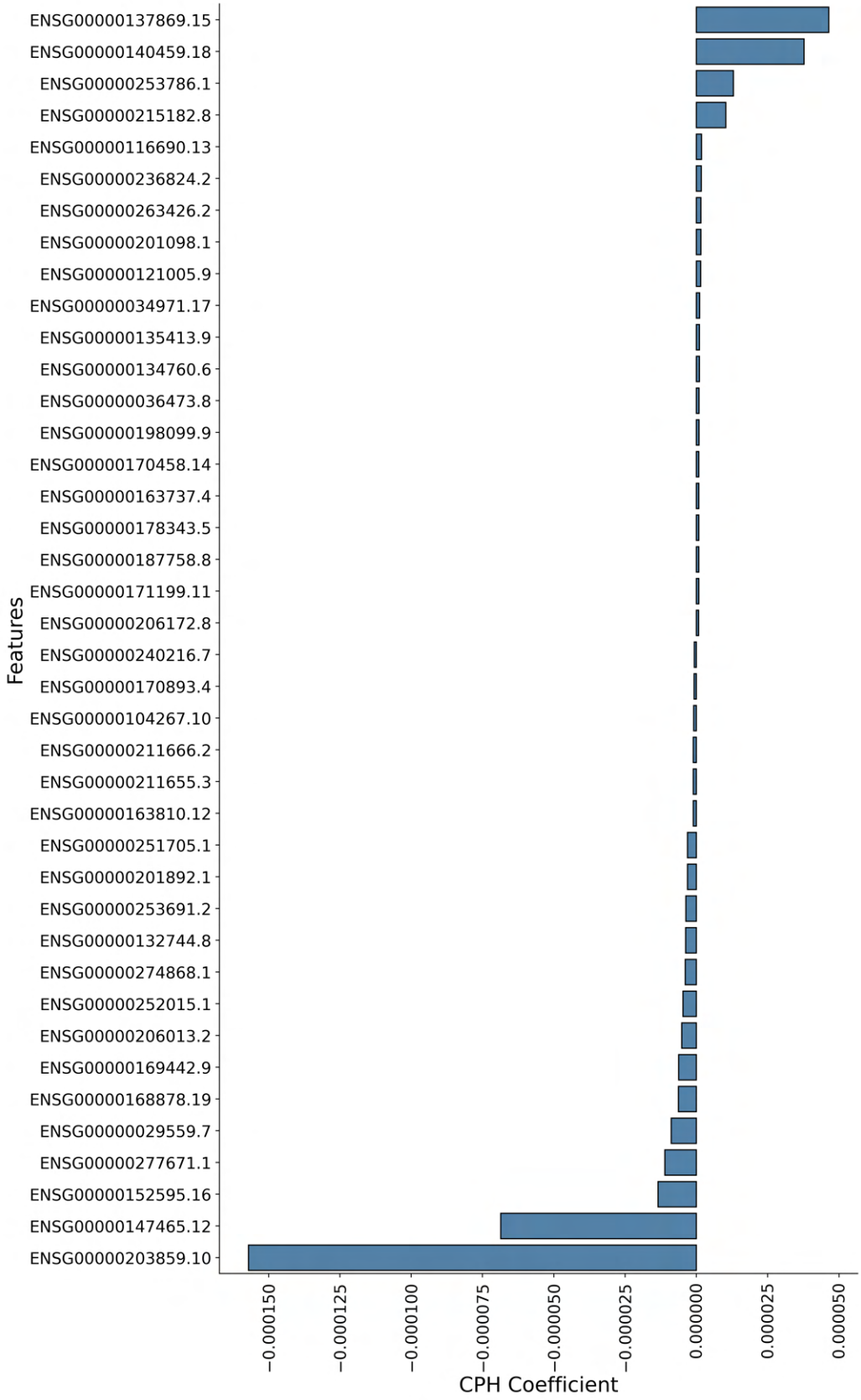

Figure 12: CPH feature coefficients for top 40 features for Run 6

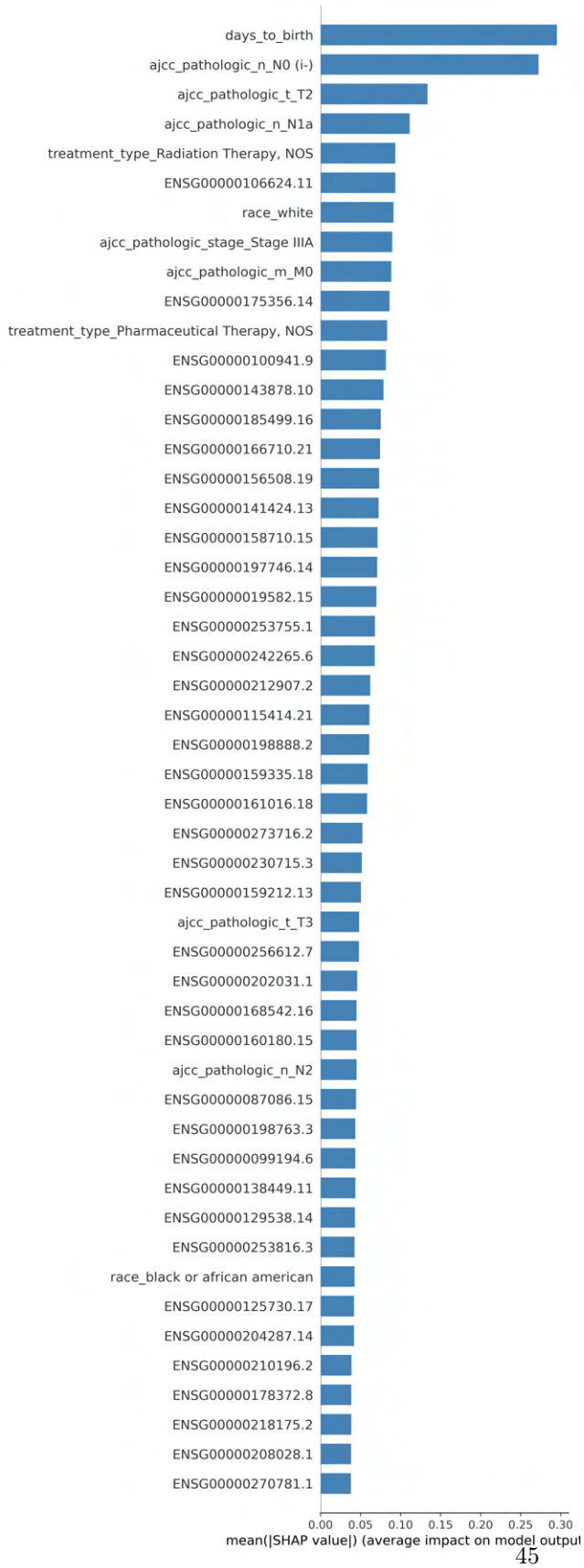

Figure 13: SHAP summary plot for DeepSurv for Run 6

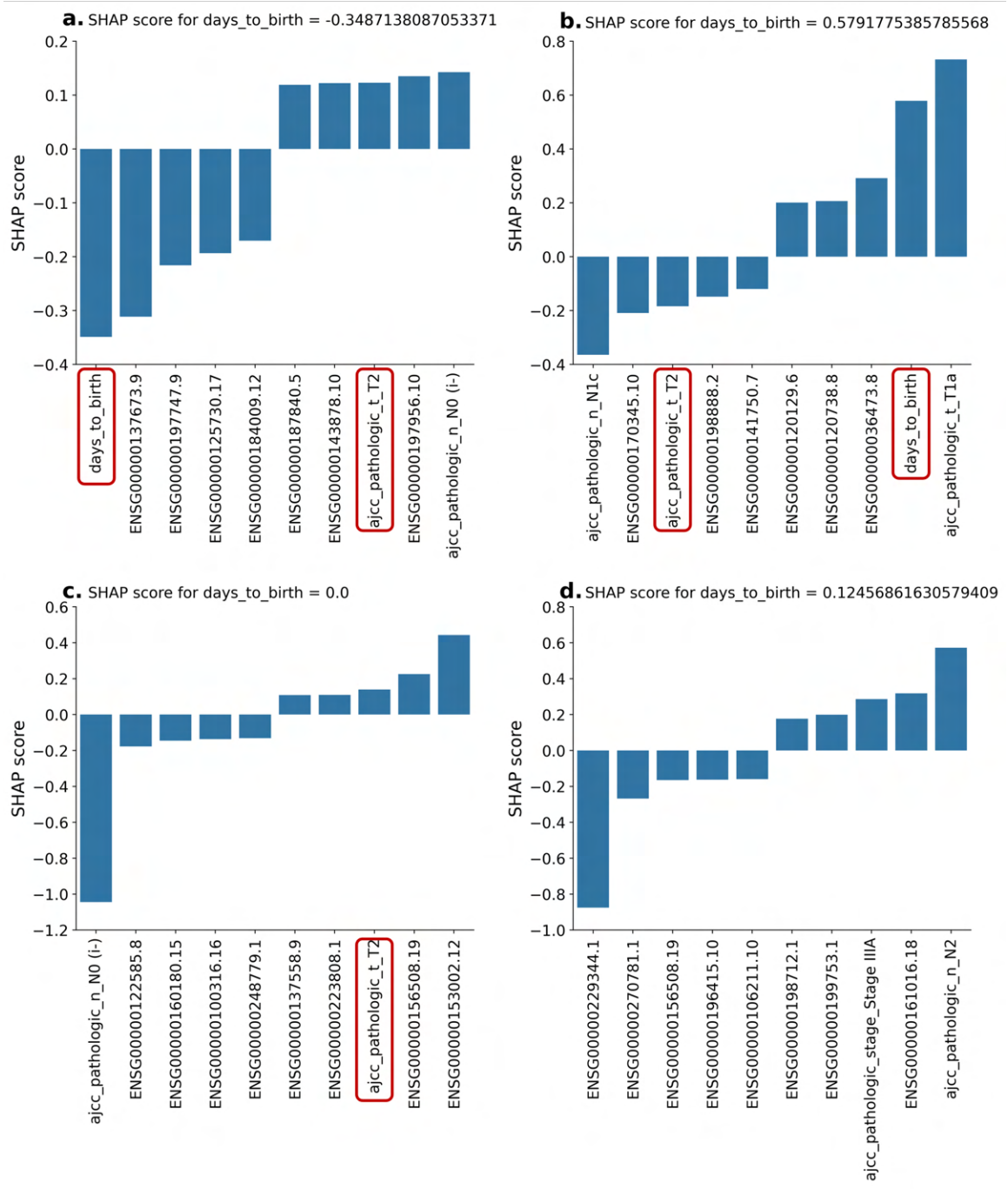

Figure 14: SHAP scores for top 10 features for four test patients for Run 6 using DeepSurv model