Bridging Structured and Unstructured Data in Stress Echocardiography Survival Analysis

Saulė Satkauskienė[1], Linas Petkevičius[1], Laura Balkevičienė[2,3] and Jelena Čeliutkienė[2,3]

[1]Vilnius University, Faculty of Mathematics and Informatics, Institute of Computer Science, Vilnius, Lithuania
[2]Clinic of Cardiac and Vascular Diseases, Faculty of Medicine, Institute of Clinical Medicine, Vilnius University, Vilnius, Lithuania
[3]Centre of Cardiology and Angiology, Vilnius University Hospital Santaros Klinikos, Vilnius, Lithuania

Abstract

The research innovatively incorporates unstructured patients diagnosis data processing using the BERT language model, subsequently analysing it with logistic regression and canonical correlation analysis to understand its relationship with structured data. This approach enriches the predictive model and underscores the value of integrating diverse data types. This processed data is then integrated alongside structured data, significantly enhancing the accuracy of the survival models. This incorporation demonstrates the potent impact of combining diverse data types on improving predictive capabilities in medical prognostics.

Keywords

Survival analysis, echocardiography data, BERT, canonical correlation, unstructured data

1. Introduction

Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide, despite substantial advances in diagnostic and therapeutic capabilities [1]. While global CVD mortality rates have demonstrated a long-term declining trend, recent evidence indicates stagnation or increases in several high-income countries [2]. Lithuania exemplifies this concerning pattern, with an age-standardised CVD mortality rate exceeding that of comparable high-income nations by more than twofold [3]. This disparity highlights the critical need for enhanced approaches to early detection, outcome prediction, and management of cardiovascular risk, particularly through advanced data-driven methodologies in echocardiography and patient monitoring systems.

Machine learning (ML) has established itself as a cornerstone technology in cardiology, with expanding applications in specialized domains such as echocardiography-based survival analysis [4]. Historically, clinical prognostication relied predominantly on ejection fraction measurements and documented comorbidities. However, contemporary research demonstrates that integrating echocardiography-derived measurements with data from electronic health records provides substantially greater predictive capacity than either source independently [5]. This integrated approach represents a meaningful shift in how researchers conceptualize the relationship between structured clinical measurements and patient outcomes.

The emergence of large language models (LLMs) has introduced novel opportunities for cardiovascular research [6]. By leveraging advanced natural language processing (NLP) capabilities, these models enable analysis of unstructured clinical text—physician notes and patient narratives — in ways previously impractical with conventional ML approaches [7]. A notable contribution in this area is the BERTSurv model proposed by Y. Zhao and colleagues [8]. This work employs BERT (Bidirectional Encoder Representations from Transformers) to extract structured features from clinical narratives. The key methodological contribution lies in combining text-derived features with structured variables such as respiratory rate and body temperature (see Figure 1), yielding improved predictive accuracy for trauma patient outcomes. This demonstrates a fundamental principle: the integration of numerical, categorical, and textual data sources yields more comprehensive clinical insight than isolated data modalities.

IVUS 2025: Information Society and University Studies, May 15, 2025, Kaunas, Lithuania
saule.satkauskiene@gmail.com (S. Satkauskienė)

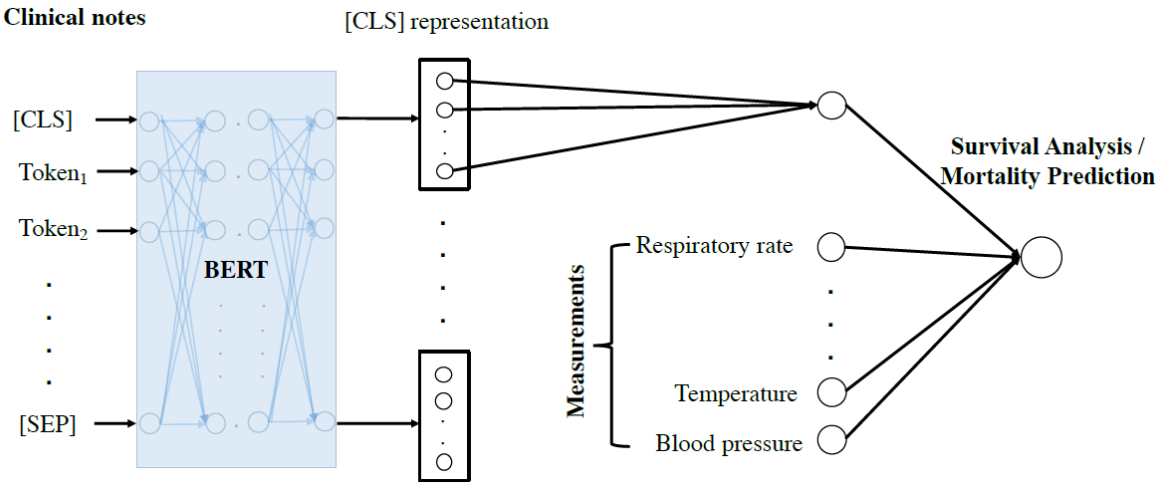

Figure 1: The framework of BERTSurv [ZHZ+21]

This paper proposes investigating how unstructured Lithuanian-language clinical notes can be effectively converted into structured data using large language models (LLMs) to support survival analysis. We employ BERT to process the clinical text, then use logistic regression and canonical correlation analysis to understand how these extracted features relate to existing structured variables. Rather than treating unstructured and structured data separately, we demonstrate that combining them substantially improves model accuracy.

2. Methodology

This section presents the methodological framework underlying the study. The approach combines processing of both structured and unstructured data, drawing on classical statistical methods alongside contemporary ML techniques.

2.1. Survival analysis

Survival analysis investigates the time until an event occurs, accounting for censoring—cases where event times are only partially observed.

The survival function (SF), denoted as $S(t)$ represents the probability of an individual surviving beyond time t:

$$S(t) = P(T > t), \quad t > 0.$$

This study focuses on right censoring. Thus, the observed time is $X = \min(T, C)$, with C as the censoring time and the indicator $\delta = 1_{T<C}$ denoting whether the event was observed ($\delta = 1$) or censored ($\delta = 0$).

Consider a sample, where each observation i is associated with a covariate vector $z^{(i)} = (z_{i1}, \ldots, z_{ip})^T$. The dataset is structured as a right-cencored sample, which can be expressed as:

$$(X_1, \delta_1, z^{(1)}), \ldots, (X_n, \delta_n, z^{(n)}).$$

Another important concept is hazard function (HF) $h(t)$, which measures the instantaneous rate of failure at time t, given survival until time t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(T \in (t, t + \Delta t] | T > t)}{\Delta t}.$$

From this definition, another important concept can be derived - the cumulative hazard function (CHF). This function accumulates the hazard over time, providing a comprehensive picture of the risk dynamics up to time t:

$$H(t) = \int_0^t h(u)du.$$

In survival analysis, three distinct types of models are identified: parametric, semi-parametric, and non-parametric. Parametric survival models presuppose that survival times are distributed according to a specific parametric distribution, exemplified by the exponential or Weibull distribution. Conversely, semi-parametric models enhance flexibility by eschewing the assumption of a predefined survival time distribution while concurrently accommodating covariates. Non-parametric models, in contrast, refrain from any presumptions regarding the survival times' distribution.

2.1.1. Weibull Accelerated Failure Time Model

In survival analysis, the Accelerated Failure Time Weibull model is revered for its robust flexibility and its adeptness in modelling a broad spectrum of survival data. At the heart of this model lies the Weibull shape parameter, ν. Additionally, one-dimensional covariates $z_1, \ldots, z_p$ are introduced, comprising the covariate vector $\mathbf{z} = (1, z_1, \ldots, z_p)^T$. In tandem, the vector of unknown parameters is denoted by $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$.

The SF is defined as:

$$S(t|\mathbf{z}) = \exp\{-\left(\frac{t}{e^{\boldsymbol{\beta}^T\mathbf{z}}}\right)^{\nu}\}.$$

2.1.2. Cox Proportional Hazards Model

The semi-parametric Cox proportional hazards (CPH) model stands out as a particularly prevalent choice. It is formulated as follows:

$$h(t|z) = h_0(t)e^{\beta z},$$

where $h_0(t)$ denotes the baseline HF, $z = (z_1, \ldots, z_p)^T$ represents the covariates, and $\beta = (\beta_1, \ldots, \beta_p)$ signifies the corresponding regression coefficients.

The key feature of this model is that it allows the estimation of the hazard ratios (HR) for different covariates without specifying the form of the baseline HF, $h_0(t)$ - this is what makes it semi-parametric.

The HR between two sets of covariates $z^{(1)}$ and $z^{(2)}$ is given by the following expression:

$$HR = \frac{h(t|z^{(1)})}{h(t|z^{(2)})} = \frac{h_0(t)e^{\beta z^{(1)}}}{h_0(t)e^{\beta z^{(2)}}} = e^{\beta(z^{(1)} - z^{(2)})}.$$

2.1.3. Random Survival Forests

Random Survival Forests (RSF) adapt the random forest framework to handle censored survival data, allowing the model to capture complex nonlinear patterns and feature interactions without requiring parametric assumptions [9]. Trees are grown on bootstrap samples and their predictions combined, which enhances both robustness and predictive accuracy.

Splitting in RSF is based on survival-specific criteria, typically maximising the log-rank statistic between daughter nodes [10]:

$$L(Z, c) = \frac{\sum_{j=1}^{m}\left(d_{j,L} - \frac{Y_{j,L}d_j}{Y_j}\right)}{\sqrt{\sum_{j=1}^{m}\frac{Y_{j,L}}{Y_j}\left(1 - \frac{Y_{j,L}}{Y_j}\right)\left(\frac{Y_j - d_j}{Y_j - 1}\right)d_j}},$$

where m is the number of distinct event times; $d_{j,L}$ and d_j are the numbers of events in the left daughter node and the parent node, respectively; and $Y_{j,L}$ and Y_j are the numbers of individuals at risk

at time t_j in those nodes. The optimal split is the covariate Z^* and threshold c^* that maximise $L(Z, c)$, leading to the greatest separation in survival between daughter nodes.

For each terminal node h, the cumulative hazard function (CHF) and SF are estimated as:

$$H_h(t) = \sum_{t_{j,h} \leq t} \frac{d_{j,h}}{Y_{j,h}}, \quad S_h(t) = \prod_{t_{j,h} \leq t} \left(1 - \frac{d_{j,h}}{Y_{j,h}}\right),$$

where $d_{j,h}$ and $Y_{j,h}$ denote, respectively, the number of deaths and individuals at risk at time $t_{j,h}$ within node h.

Each bootstrap sample used to grow the survival trees contains approximately two-thirds of the observations, leaving one-third out. The Out-of-bag (OOB) observations can be used to test the tree, as they were not seen during the tree's construction.

Let $I_i \in \{0, 1\}$ denote whether case i belongs to OOB sample, with $I_i = 1$ corresponding to OBB cases. The OOB estimators for the CHF and the SF are then given by:

$$H^{OOB}(t|Z_i) = H_h(t), \quad S^{OOB}(t|Z_i) = S_h(t), \text{ if } Z_i \in h \text{ and } I_i = 1.$$

A key advantage of RSFs is their nonparametric approach to variable importance (VIMP). The method centers on how much prediction accuracy declines when a variable is removed or permuted, offering a practical measure of each variable's predictive value.

VIMP for a variable Z is computed as follows:

1. **Random Permutation of OOB Values:** For each tree in the forest, randomly shuffle the out-of-bag (OOB) values of Z while leaving all other variables unchanged. This step disrupts the relationship between Z and the response, isolating its effect on predictive accuracy.
2. **Propagate Permuted OOB Data:** Pass the permuted OOB data down the tree to reassess the tree's prediction accuracy. Calculate the new OOB error for each tree after permutation.
3. **Calculate Increase in Error:** Determine the increase in prediction error for each tree due to the permutation of Z. This increase is a measure of the importance of Z in that particular tree.
4. **Average Increase Across Trees:** Average the increases in prediction error across all trees in the forest to derive the VIMP for Z. A higher VIMP value indicates a greater importance of Z in predicting the outcome.

This procedure isolates the impact of the variable Z on the model's predictive accuracy, quantifying its importance through the increase in prediction error caused by its permutation.

2.1.4. Models comparison

The concordance index (C-index) is an evaluative measure that does not require a predetermined time point for assessing the performance of a model; rather, it incorporates the censoring status of individuals [11]. Computing the C-index involves the following procedure [10]:

1. Generate all possible pairwise comparisons across the entire dataset.
2. Exclude pairs where the shorter event time is censored. Additionally, discard pairs (i, j) if $X_i = X_j$, except when $(\delta_i = 1, \delta_j = 0)$ or $(\delta_i = 0, \delta_j = 1)$. Denote the remaining pairs as $\mathbb{S}$, with $permissible = |\mathbb{S}|$.
3. When $X_i \neq X_j$, assign 1 point to each $s \in \mathbb{S}$ where the observation with the shorter time also receives a worse predicted survival estimate.
4. When $X_i \neq X_j$, assign 0.5 points to each $s \in \mathbb{S}$ where model predictions match.
5. When $X_i = X_j$, assign 1 point to each $s \in \mathbb{S}$ where predictions coincide.
6. When $X_i = X_j$, assign 0.5 points to each $s \in \mathbb{S}$ where predictions diverge.
7. Define $concordance$ as the total count for all permissible pairs, then compute the C-index as

$$Cindex = \frac{\text{concordance}}{\text{permissible}}$$

Table 1
Top 20 structured variables identified through logistic regression on BERT-Processed textual data.

Variable	Accuracy	Balanced accuracy	Specificity	Sensitivity	PPV	NPV	Kappa	p-value
Test.performed.by_8	0.94	0.97	0.98	0.91	0.98	0.89	0.88	0.00
Conclusion	0.93	0.95	0.97	0.89	0.97	0.89	0.86	0.00
Test.performed.by_10	0.92	0.97	0.98	0.86	0.99	0.68	0.74	0.00
Visual.assessment_-	0.91	0.95	0.86	0.97	0.87	0.97	0.83	0.00
Test.performed.by_1	0.90	0.95	0.97	0.83	0.96	0.89	0.82	0.00
Previous.myocardial.infarction	0.90	0.93	0.84	0.96	0.89	0.95	0.82	0.00
Test.performed.by_6	0.89	0.96	0.98	0.80	0.97	0.86	0.81	0.00
Test.performed.by_11	0.88	0.98	0.98	0.78	0.99	0.68	0.71	0.00
Test.performed.by_5	0.88	0.97	0.99	0.78	0.98	0.82	0.78	0.00
Visual.assessment_+	0.88	0.94	0.97	0.79	0.96	0.83	0.78	0.00
Test.performed.by_13	0.87	0.98	0.99	0.76	0.99	0.62	0.68	0.00
Tension.in.the.chest_0	0.87	0.88	0.83	0.91	0.88	0.88	0.75	0.00
Tension.in.the.chest_3	0.87	0.88	0.91	0.83	0.88	0.88	0.75	0.00
PTCA.atherectomy	0.87	0.92	0.78	0.96	0.85	0.94	0.76	0.00
Coronary.heart.disease	0.86	0.87	0.81	0.91	0.83	0.90	0.72	0.00
Diagnosis_Diagnostic	0.86	0.86	0.82	0.89	0.87	0.85	0.72	0.00
Test.performed.by_7	0.85	0.97	0.99	0.72	0.98	0.81	0.75	0.00
Diagnosis_CHD	0.85	0.86	0.90	0.80	0.86	0.86	0.71	0.00
CABG	0.85	0.95	0.73	0.98	0.81	0.96	0.74	0.00
Test.performed.by_2	0.84	0.96	0.99	0.70	0.97	0.85	0.74	0.00

8. Compute the prediction error as $Error = 1 - Cindex$, where $0 \leq Error \leq 1$. A C-index of 1 (error of 0) signals perfect discrimination, while $Cindex = Error = 0.5$ suggests performance no better than random chance.

2.2. Unstructured data analysis

LLMs are reshaping NLP by providing advanced methodologies for effectively analyzing and extracting insights from abundant but unstructured data. BERT is a prime example of an advanced NLP model that has been engineered to process and understand language in a nuanced manner [12]. Unlike earlier models, BERT operates bidirectionally, meaning it analyzes text by looking at the words that come before and after a given word [13]. This capability is crucial when dealing with complex texts, such as clinical notes, which often contain intricate expressions and specialized terminology [14].

2.2.1. Unstructured Data Processing

A text collection $D = {T^{(i)}}_{i=1}^{N}$ was processed using the *bert-base-multilingual-uncased* model, which supports multiple languages, including Lithuanian. Each text $T^{(i)}$ was tokenised using the WordPiece method, producing a sequence $S^{(i)} = t_1^{(i)}, t_2^{(i)}, \ldots, t_k^{(i)}$.

Tokenised sequences were standardised to a fixed maximum length K through padding, resulting in $\tilde{S}^{(i)}$. Each token $t_j^{(i)}$ was then mapped to an integer identifier from the model's predefined vocabulary of 105,879 tokens, yielding the numerical sequence $\hat{S}^{(i)}$.

The padded sequences $\hat{S}^{(i)}$ were passed through the BERT model $\mathcal{M}$, generating contextual embeddings represented by hidden states $H^{(i)}$. Each token embedding is a 768-dimensional vector capturing the semantic and syntactic context of the text.

2.2.2. Logistic Regression

Logistic regression was applied to interpret BERT-derived features and assess their association with binary clinical outcomes. The model estimates the probability of an event as

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_d X_d)}},$$

where $X_1, \ldots, X_d$ are predictors and $\beta_0, \ldots, \beta_d$ are coefficients optimised through maximum likelihood estimation.

Model performance was evaluated using several classification metrics:

$$\text{Accuracy} = \frac{TP + TN}{N}, \quad \text{Sensitivity} = \frac{TP}{TP + FN},$$

$$\text{Specificity} = \frac{TN}{TN + FP},$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2},$$

$$\text{PPV} = \frac{TP}{TP + FP}, \quad \text{NPV} = \frac{TN}{TN + FN}.$$

Cohen's Kappa (κ) was used to assess the level of agreement between predicted and true outcomes while accounting for chance agreement. It captures this by comparing observed to expected agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o represents the proportion of cases where predictions matched true labels, and p_e is the proportion expected if predictions were random. A kappa of 1 signals perfect agreement, whereas 0 suggests the model performs no better than random chance.

Statistical significance of regression coefficients was assessed through p-values, which quantify the probability that observed associations between predictors and outcomes could have arisen through chance.

2.2.3. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) identifies linear relationships between two multivariate variable sets, $\mathbf{X}$ and $\mathbf{Y}$ [15]. The method determines the weight vectors $\mathbf{a}$ and $\mathbf{b}$ that define linear combinations $U = \mathbf{a}^T\mathbf{X}$ and $V = \mathbf{b}^T\mathbf{Y}$, maximising the correlation between them:

$$\text{Cor}(U, V) = \frac{\mathbf{a}^T\mathbf{\Sigma}_{XY}\mathbf{b}}{\sqrt{\mathbf{a}^T\mathbf{\Sigma}_X\mathbf{a}}\sqrt{\mathbf{b}^T\mathbf{\Sigma}_Y\mathbf{b}}}$$

subject to unit variance of U and V. This procedure yields $\min(p, q)$ canonical variate pairs (U_k, V_k), where the k-th canonical correlation coefficient is $\rho_k = \text{Cor}(U_k, V_k)$.

2.2.4. Principal Components Analysis

Principal Components Analysis (PCA) is an orthogonal transformation that converts correlated variables into linearly uncorrelated Principal Components (PCs) for dimensional reduction.

Given a dataset $\mathbf{A} \in \mathbb{R}^{n \times p}$ with n observations and p variables, PCA transforms the data such that the first PC has maximum variance, and each subsequent component maximizes variance while remaining orthogonal to previous components. The k-th principal component is defined as:

$$PC_k = \mathbf{a}_k^T A$$

where $\mathbf{a}_k$ is the k-th eigenvector of the covariance matrix $\mathbf{\Sigma}_A$. Components are selected based on the proportion of variance retained, with the first few PCs typically capturing the majority of dataset variance while reducing dimensionality.

3. Experiments

3.1. Database and data preparation

The dataset[1] covers the period from January 2002 to June 2022 and was later enhanced with additional mortality data up to 13 July 2023. After manual cleaning and removal of irrelevant records and variables, the dataset comprised 14.5K entries and 63 variables.

3.2. Data Processing

The study applies a BERT-based text embedding model to analyse unstructured textual data. Each text is tokenised, padded, and numerically encoded before being processed by the BERT model, which generates hidden state vectors $H^{(i)}$ in a 768-dimensional space. These hidden states capture contextual and linguistic information of the text. To obtain a concise representation, the mean of the token embeddings is computed:

$$\bar{H}^{(i)} = \frac{1}{K} \sum_{j=1}^{K} H_j^{(i)},$$

This yields a single vector per sequence, providing an interpretable summary of each text's semantic content.

3.2.1. BERT-Processed Textual Data Analysis by Logistic Regression

Logistic regression was applied to the BERT-derived text embeddings to evaluate their predictive value and identify the most informative features. The model examined the relationship between these features and the target outcome, such as diagnostic or clinical status.

The top 20 most significant variables, ranked by balanced accuracy, are shown in Table 1. Variables were colour-coded to distinguish clinically relevant factors (green) from non-clinical ones (pink).

The analysis revealed that predictive performance stemmed not only from clinical indicators but also from stylistic variations in medical documentation. Non-clinical linguistic patterns, as reflected in the pink variables, contributed significantly to model predictions, alongside clinically meaningful variables such as diagnostic terms and procedural references. This finding demonstrates that the model extracts both clinically relevant content and documentation style from unstructured medical texts, offering valuable information for integration with structured data in predictive and survival analyses.

3.2.2. Interpretation of Canonical Correlation Analysis Between Structured and Unstructured Data

Canonical Correlation Analysis (CCA) was used to explore associations between structured clinical variables, represented as dummy-coded predictors $\mathbf{X} = (X_1, \ldots, X_{123})$, and BERT-derived unstructured text embeddings $\mathbf{Y} = (Y_1, \ldots, Y_{768})$.

The analysis yielded 123 canonical correlations ρ_1 ranging from 0.93 to 0.12, with the first canonical correlation $\rho_1 = 0.93$ indicating a strong linear relationship between the two data modalities. Subsequent correlations, though smaller, remained notable (e.g., 0.89, 0.88), demonstrating consistent shared variance across several canonical variate pairs.

Correlation analysis between $\mathbf{X}$ and the first canonical variate of $\mathbf{Y}$ revealed that variables such as *Conclusion* (0.77) and *Visual Assessment* (0.72) were most strongly associated with the unstructured text component (Table 2). Additional myocardial segment indicators also displayed moderate correlations, reinforcing the relevance of linking structured clinical parameters with unstructured medical notes.

Statistical tests (Wilks' Lambda, Hotelling–Lawley Trace, and Pillai–Bartlett Trace) confirmed that the first 80–90 canonical correlations were significant ($p < 0.05$), suggesting meaningful relationships

[1]Lithuanian Bioethics permission -AI (2023-03-25) granted to the PI at Vilnius University Hospital Santaros klinikos.

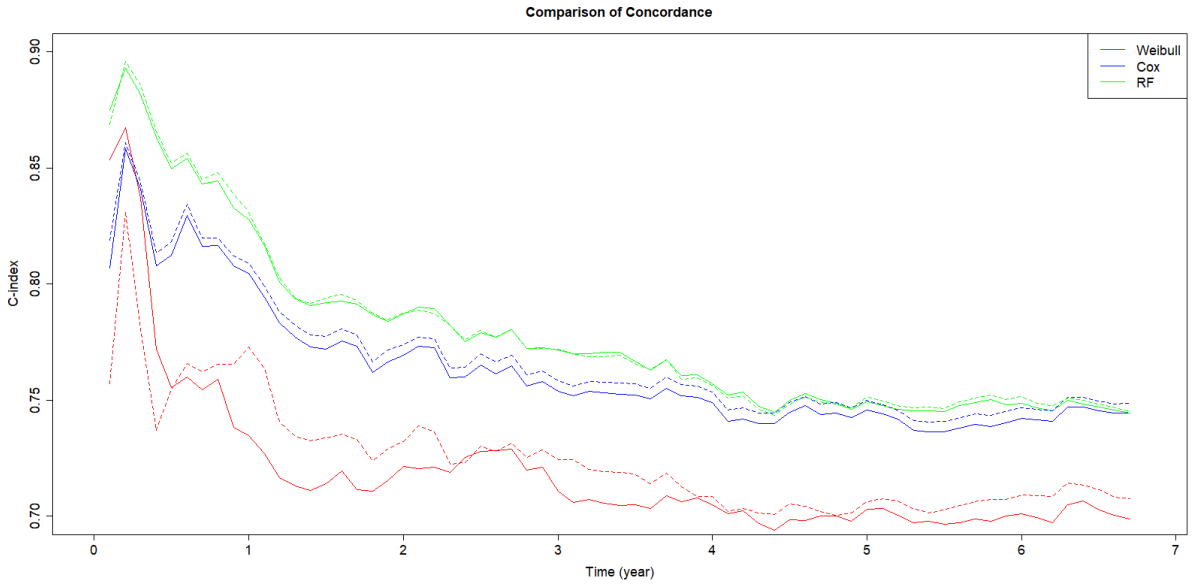

Figure 2: Temporal evolution of concordance index for Weibull, Cox Proportional Hazards, and Random Survival Forests models.

Note: Solid lines represent structured data models; dashed lines include unstructured data.

between structured and unstructured representations. Beyond this range, p-values exceeded 0.05, indicating diminishing association strength.

3.3. Structured and Unstructured Data Survival Analysis

Findings from logistic regression and CCA indicated that BERT-processed clinical notes contain valuable information. The variable *Conclusion*, excluded from structured survival analysis due to missing values, appeared well represented in the text embeddings. To assess its potential contribution, survival analysis was condcted on structured data and on mixed data (structured and unstructured) by integrating the first two principal components (PCs) of the BERT-processed notes. The number of PCs was limited for computational efficiency and interpretability.

The concordance index (c-index) plot (Figure 2) suggests minor yet consistent gains in model performance when clinical notes are included. The CPH and RSF models show improved concordance across most time intervals, particularly between 4.5 and 5 years, where BERT-derived features significantly enhance the CPH model. Towards the end of the follow-up, its performance approaches or surpasses that of RSF.

In the short term (up to six months), unstructured data slightly reduced the Weibull model's c-index. From 0.5 to 2.4 years, however, its performance improved notably, followed by moderate gains beyond 2.6 years. Overall, BERT-processed clinical notes appear to narrow performance differences between CPH and RSF models during long-term follow-up. Considering its simplicity and computational efficiency, the CPH model may be preferable for extended prognoses, while RSF retains superiority in early outcome prediction.

4. Conclusions

In the analysis of cardiovascular prognosis models, the RSF algorithm demonstrated the highest predictive performance, as reflected by its superior c-index. This finding highlights the relevance of advanced ML methods in refining prognostic accuracy within cardiology. Furthermore, the high canonical correlation obtained using the BERT model to process unstructured clinical text underscores the added value such data can provide when effectively integrated with structured datasets. Incorporating

Table 2
Top 20 correlations between structured data and first canonical variate of BERT-processed unstructured data.

Variable	Correlation
Conclusion	0.77
Visual.assessment_+	0.72
Middle.inferior.segment.assessment	0.49
Basal.inferior.segment	0.47
Tension.in.the.chest	0.46
Ejection.fraction.after.stress	-0.45
Previous.myocardial.infarction	0.45
CHD	0.43
Diagnosis_CHD	0.42
Chest.pain	0.42
Pain_+	0.39
Basal.septal.inferior.segment	0.38
Upper.lower.segment	0.37
Median.inferior.lateral.segment	0.36
Median.septal.inferior.segment	0.35
Upper.septal.segment	0.33
Basal.inferior.segment	0.33
Basal.inferior.lateral.segment	0.33
Reason.for.conducting.the.test_Prognostic	0.33
Middle.inferior.segment	0.32

BERT-processed unstructured data into survival analysis models notably improved their predictive capability, supporting a comprehensive modelling approach that combines multiple data modalities to achieve more robust and informative predictions.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] W. H. Organization, The top 10 causes of death, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2020.

[2] G. A. Roth, G. A. et al, Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the gbd 2019 study, Journal of the American College of Cardiology 76 (2020) 2982–3021. URL:

https://www.sciencedirect.com/science/article/pii/S0735109720377755. doi:https://doi.org/10.1016/j.jacc.2020.11.010.

[3] T. I. for Health Metrics, Evaluation, Global health data exchange, ???? URL: http://ghdx.healthdata.org/gbd-results-tool.

[4] Y. Arfat, G. Mittone, R. Esposito, B. Cantalupo, G. Ferrari, M. Aldinucci, A review of machine learning for cardiology, Minerva Cardioangiologica (2021). doi:10.23736/S2724-5683.21.05709-4.

[5] M. D. Samad, A. Ulloa, G. J. Wehner, L. Jing, D. Hartzel, C. W. Good, B. A. Williams, C. M. Haggerty, B. K. Fornwalt, Predicting survival from large echocardiography and electronic health record datasets: Optimization with machine learning, JACC: Cardiovascular Imaging 12 (2019) 681–689. URL: https://www.sciencedirect.com/science/article/pii/S1936878X18303851. doi:https://doi.org/10.1016/j.jcmg.2018.04.026.

[6] J. Clusmann, F. Kolbinger, H. Muti, Z. Carrero, J.-N. Eckardt, N. Laleh, C. Löffler, S.-C.Schwarzkopf, M. Unger, G. Veldhuizen, S. Wagner, J. Kather, The future landscape of large language models in medicine, Communications medicine 3 (2023) 141. doi:10.1038/s43856-023-00370-1.

[7] M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. Pfeffer, J. Fries, N. Shah, The shaky foundations of large language models and foundation models for electronic health records, npj Digital Medicine 6 (2023). doi:10.1038/s41746-023-00879-8.

[8] Y. Zhao, Q. Hong, X. Zhang, Y. Deng, Y. Wang, L. R. Petzold, Bertsurv: Bert-based survival models for predicting outcomes of trauma patients, CoRR abs/2103.10928 (2021). URL: https://arxiv.org/abs/2103.10928. arXiv:2103.10928.

[9] I. Bou-Hamad, D. Larocque, H. Ben-Ameur, A review of survival trees, Statistics Surveys 5 (2011). doi:10.1214/09-SS047.

[10] H. Ishwaran, M. S. Lauer, E. H. Blackstone, M. Lu, U. B. Kogalur, randomForestSRC: random survival forests vignette, http://randomforestsrc.org/articles/survival.html, 2021. URL: http://randomforestsrc.org/articles/survival.html.

[11] B. Weathers, D. R. Cutler, Comparison of survival curves between cox proportional hazards, random forests, and conditional inference forests in survival analysis, 2017. URL: https://digitalcommons.usu.edu/gradreports/927/.

[12] J. Devlin, M. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. doi:arXiv:1810.04805.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[14] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, npj Digital Medicine 4 (2021) 86. doi:10.1038/s41746-021-00455-y.

[15] N. Helwig, Canonical correlation analysis, PowerPoint slides, 2017. URL: http://users.stat.umn.edu/~helwig/notes/cancor-Notes.pdf, university of Minnesota.