Predicting Sepsis Onset with Survival Analysis Over Signature Transformations

Research Paper

James P. Howard, II[*]
Whiting School of Engineering
The Johns Hopkins University
Baltimore, Maryland, United States
james.howard@jhu.edu

Thomas B. Woolf
School of Medicine, Department of Physiology
The Johns Hopkins University
Baltimore, Maryland, United States
twoolf@jhu.edu

May 27, 2021

Abstract

Predicting clinical outcomes from time-series medical data is a complex but essential endeavor. In this study, we propose a novel approach that combines traditional survival models like Cox proportional hazards, logistic, and multi-task logistic regression (MTLR) with the robust mathematical framework of signature methods. These methods are particularly effective in capturing the underlying dynamics of time-series data with stochastic error. We introduce the concept of rough paths to provide a foundational understanding of how these techniques can capture not only the data's deterministic aspects but also its stochastic nature, thereby enriching the feature set used for making more accurate predictions. Our methodological pipeline consists of feature generation, the application of signature methods, and the implementation of a survival analysis neural network. We demonstrate that our approach is computationally efficient, eliminating the need for graphics processing unit (GPU)-intensive calculations and thereby making it more accessible for real-world applications. Results show a promising increase in predictive power. This study sets the groundwork for future research in harnessing the mathematical sophistication of signature methods for more nuanced and accurate predictions in the medical field.

Keywords survival analysis · signature methods · time series analysis

1 Introduction

Sepsis, a systemic inflammatory response to infection, stands as one of the most formidable challenges faced by healthcare professionals today. This potentially life-threatening condition is precipitated by various pathogens, with bacterial infections being the most prevalent. Symptoms are diverse and encompass fever, increased heart rate, rapid breathing, and confusion. As the disease progresses, the symptoms can escalate to severe sepsis, marked by organ dysfunction, and culminate in septic shock, where the body's blood pressure drops critically, resistant to standard fluid resuscitation [Hunt, 2022].

The early detection and prompt treatment of sepsis are imperative in determining patient outcomes. Delays in identifying and managing sepsis can quickly lead to organ failure and increase the risk of mortality. In the realm of healthcare, there is a growing recognition of the need for accurate and timely diagnostic tools to pinpoint patients at risk of sepsis or those in the early stages. The necessity is twofold: firstly, to administer the right treatment rapidly, and secondly, to optimize the allocation of often limited medical resources.

Existing prediction models have ventured into this domain, offering insights and tools for the early identification of sepsis. However, while some have shown promise, there remain gaps in both diagnostic and prognostic capabilities,

[*] Website: https://jameshoward.us

with issues such as false positives or late detections still prevalent. Our endeavor, in this context, is to develop or validate a multivariable prediction model that not only addresses these limitations but also integrates seamlessly into clinical workflows, facilitating real-time decision-making and improving patient outcomes.

The primary objective of our study is to develop an advanced predictive model tailored to forecast the onset of sepsis using comprehensive clinical data. Sepsis, a complex systemic inflammatory response, presents diagnostic and treatment challenges within the medical community. Accurate and timely prediction of its occurrence can profoundly influence patient outcomes, ensuring rapid intervention and optimal allocation of medical resources. Our study aims to bridge existing gaps in early detection by harnessing the capabilities of the signatory method [Kidger and Lyons, 2020] in conjunction with survival analysis techniques implemented in PyTorch.

The impetus for focusing solely on model development, as opposed to validation stems from the innovative nature of our approach. Integrating the signatory method with survival methods such as Cox regression [Kvamme et al., 2019], Cox proportional hazards [Katzman et al., 2018], PC-Hazard [Kvamme and Borgan, 2021], partial logistic regression [Biganzoli et al., 1998], and multi-task logistic regression (MTLR) [Yu et al., 2011, Fotso, 2018], we aim to design a model with enhanced accuracy and predictive power. The comprehensive nature of the sepsis dataset, encompassing vital signs, laboratory values, and demographic information provides an unparalleled opportunity to discern intricate patterns and dependencies often elusive in traditional models. Following the successful development of this model, subsequent research will venture into its rigorous validation, thereby ascertaining its efficacy and applicability in real-world clinical settings.

1.1 Applications of Rough Paths

Rough paths theory is a branch of mathematics that has garnered significant attention for its ability to provide a rigorous framework for understanding the behavior of controlled differential equations, and more broadly, stochastic processes. Originating from the work of Terry Lyons in the late 1990s [Lyons, 1998, Lyons and Qian, 1997], the formal mathematical definition of a rough path $\mathbb{P}$ involves a pair of continuous paths $(X, \mathbb{X})$, where $X : [0, T] \to \mathbb{R}^d$ is the original path and $\mathbb{X}$ is its so-called "lift," capturing the area under the path to provide a second-order description of its behavior. Together, these constructs provide an enriched structure that enables more effective integration against irregular signals, or more formally, against paths lacking smoothness.

In machine learning, rough paths have been used to extend the utility of recurrent neural networks for tasks such as handwriting recognition [Yang et al., 2016]. Their significance is not merely theoretical; rough paths offer practical advantages for numerical approximations and simulations of complex systems, where traditional methods may not yield satisfactory results. One of the most salient aspects of rough paths is their capacity to serve as a powerful analytical tool for time-series data, especially those with inherent stochastic errors. In stochastic calculus, rough paths offer a way to bypass some of the limitations associated with the Ito calculus, such as issues with pathwise uniqueness or existence of solutions to stochastic differential equations. They provide an alternative, yet rigorous, mathematical foundation for understanding complex behaviors often encountered in time-series data.

Time-series data often contains inherent stochastic errors that make it difficult to model and analyze through conventional methods. Conventional models like autoregressive integrated moving average (ARIMA) and exponential smoothing fail to capture the intricate dynamics effectively [Tong et al., 2022]. This is where rough paths come into play. By converting time-series data into a rough path, we introduce an additional layer of structure, specifically designed to deal with stochastic errors [Bayer et al., 2023]. This enriched structure allows for more effective modeling of complex interactions and stochastic behaviors within the data, which are often not linear or Gaussian in nature.

The predictive value of rough paths in time-series analysis is compelling. Rough paths offer a robust way to capture the historical information embedded in time-series data, thus yielding predictive models that are both accurate and interpretable. Given that rough paths can successfully model complex interactions and capture second-order behavior, it opens up the possibility for improved forecasting. For example, in financial time-series, where rapid changes and jumps are common, rough paths can capture the intricate movements and trends that other models may miss, making them an effective tool for predicting future stock prices or economic indicators.

In machine learning, rough paths have been integrated into models like recurrent neural networks to improve their predictive power. Combining machine learning models with rough paths also allows for uncertainty quantification, which is often a critical component in real-world decision-making scenarios. Moreover, rough paths have the potential to serve as a universal feature extractor for time-series data, which can be highly beneficial for various predictive analytics tasks, especially in health care monitoring.

Signature methods and rough paths are naturally complementary, offering a robust framework for analyzing complex, high-dimensional time series data. Rough paths theory generalizes classical calculus to handle paths with stochastic noise, making it ideally suited for real-world time series that often contain erratic fluctuations. Signature methods, by virtue of computing iterated integrals over a given path, provide a compact yet expressive summary of the path [Chevyrev and Oberhauser, 2018]. When applied to rough paths, the signature encapsulates not just the geometric features of the path but also its stochastic characteristics, effectively capturing the underlying dynamics of the time series. This joint framework of rough paths and signature methods allows us to perform sophisticated analyses on noisy, high-dimensional data, providing predictive and explanatory value that is crucial in fields such as finance, healthcare, and beyond.

1.2 Signature Methods

Signature methods provide a robust mathematical framework to analyze sequential or path-like data. They are rooted in the theory of rough paths, which allows for a detailed representation of a path by capturing the infinite collection of iterated integrals [Chevyrev and Kormilitzin, 2016]. For a given path $\mathbf{X} : [a, b] \to \mathbb{R}^d$, the iterated integrals encapsulate not only the pointwise information but also the higher-order interactions between the components of the path, capturing the multi-level interactions.

The iterated integrals of a path are defined through the expression:

$$\langle \mathbf{X}^{i_1}, \ldots, \mathbf{X}^{i_k} \rangle = \int_{a<t_1<\ldots<t_k<b} d\mathbf{X}_{t_1}^{i_1} \ldots d\mathbf{X}_{t_k}^{i_k}, \tag{1}$$

Here, the notation $\langle \cdot \rangle$ represents the k-fold iterated integral, and the indices $i_1, \ldots, i_k$ denote the dimensions of the path. This expression is integrated over all possible time partitions, thereby capturing the complex interplay between different segments of the path over multiple time scales.

These iterated integrals form the building blocks of the signature of a path, which has been successfully applied in various domains such as finance, machine learning, and data analysis. By capturing the path's information up to all possible orders, they provide a rich and highly informative feature set. The signature constructed from iterated integrals benefits from mathematical properties such as translation invariance and the ability to handle highly irregular paths, making it a versatile tool for many applications.

The signature of a path is formally defined as the collection of these iterated integrals, and can be expressed as:

$$S(\mathbf{X}) = (1, \langle \mathbf{X} \rangle, \langle \mathbf{X}, \mathbf{X} \rangle, \langle \mathbf{X}, \mathbf{X}, \mathbf{X} \rangle, \ldots), \tag{2}$$

where $\mathbf{X}$ represents the path, and $\langle \cdot \rangle$ denotes the iterated Stratonovich integral.

In the context of machine learning and data analysis, signature methods are utilized to transform sequential data into a feature set that encapsulates not only the information about the path itself but also about its higher-order interactions. This transformation is often implemented using a truncated version of the signature, keeping only up to n-th level terms, and it offers invariance under reparameterizations of the path. This property makes the signature particularly robust to common variations in real-world data.

The computation of signatures can be efficiently performed using specialized algorithms, allowing for scalability in large datasets. In practice, signature features can be fed into various learning algorithms, such as support vector machines, neural networks, or gradient boosting machines, to build predictive models. The versatility and mathematical foundation of signature methods make them suitable for a wide range of applications including financial time series forecasting, human activity recognition, and many others.

By utilizing the unique properties of signatures, the model can extract complex dependencies and relationships within the data. This often results in enhanced predictive accuracy and interpretability compared to traditional methods. The integration of signature methods within the study ensures a principled approach to analyzing the intricate structures inherent in the data under examination.

1.3 Survival Analysis

Survival analysis, also known as time-to-event analysis, is a statistical approach used to predict the time until a particular event or endpoint, such as death, relapse, or failure, occurs. This methodology has broad applications across various fields such as medicine, engineering, economics, and social sciences. What sets survival analysis apart from standard regression techniques is its ability to handle censored data, where the exact time of the event is unknown or has not yet occurred.

Cox regression, is a seminal technique in survival analysis that models the hazard rate as a function of several predictors. Unlike traditional regression models, the Cox model does not assume a specific underlying distribution for the survival times. Instead, it operates on the hazard function, describing how the hazard changes in response to the covariates. Mathematically, the model is expressed as $h(t|X) = h_0(t)\exp(\beta' X)$, where $h(t|X)$ is the hazard function at time t given covariates X, $h_0(t)$ is the baseline hazard, and β represents the coefficients that quantify the effect of covariates. This formulation allows for an understanding of how different factors influence the risk of the event occurring at any given time point, adjusting for all other variables in the model. The proportionality assumption, which implies that the effects of predictors are constant over time, is central to this model and should be assessed when applying Cox regression. Its flexibility and ability to incorporate censored data have made the Cox model a cornerstone in time-to-event analysis across various fields, including medical research, economics, and engineering.

The Cox proportional hazards model extends the concept of survival analysis by allowing the incorporation of covariates that may influence the survival time. This semi-parametric model doesn't assume a specific distribution for the baseline hazard function, providing flexibility in modeling various survival data. The model is expressed as:

$$h(t|X) = h_0(t)\exp(\beta' X)$$

where $h(t|X)$ is the hazard function at time t given covariates X, $h_0(t)$ is the baseline hazard, and β represents the coefficients that quantify the effect of covariates.

The partially conditional hazard (PC-Hazard) model is a more recent addition to the family of survival analysis models. Unlike the Cox regression model, which assumes that the effects of the predictors are constant over time (proportional hazards assumption), the PC-Hazard allows for a more flexible approach. The PC-Hazard model can accommodate non-proportional hazards by introducing additional functions of time into the model, known as time-varying covariates or time-varying effects. This allows the effect of a variable to change over time, providing a richer understanding of how predictors interact with survival. This can be particularly important in studies where the relationship between covariates and the hazard function may change during the follow-up period. The inclusion of time-dependent effects can lead to a more nuanced and accurate representation of complex survival data. Consequently, the PC-Hazard model has seen applications in various scientific fields where the assumptions of the traditional Cox model may be too restrictive, offering enhanced flexibility in capturing the underlying dynamics of time-to-event data.

The logistic model in this context represents another approach to survival analysis, specifically for binary outcome prediction. Unlike traditional survival models that focus on predicting the time until an event occurs, the logistic model in this context is geared toward predicting whether an event (*e.g.*, failure, disease occurrence) happens within a specified time frame. In the logistic model, the relationship between the predictors and the log odds of the event is modeled using a logistic function. This provides the probability of the event occurring within the defined period. One of the significant advantages of using the logistic model is its simplicity and ease of interpretation. The estimated parameters can be easily transformed into odds ratios, providing a clear understanding of the effect of each predictor on the likelihood of the event.

In comparison to continuous-time survival models like the Cox regression, the logistic model treats survival time as a binary outcome (e.g., whether an event occurs within 96 hours), making it particularly suitable for scenarios where the interest lies in predicting short-term risk or categorizing individuals based on a critical time threshold. It's a valuable tool in medical studies, engineering reliability analysis, and other fields where binary classification of survival outcomes is relevant. The logistic model's applicability in this study demonstrates the broad array of methodologies available for survival analysis, each tailored to specific types of data and research questions.

The MTLR model is a distinctive approach to survival analysis that incorporates aspects of both logistic regression and continuous-time modeling. Unlike traditional survival models that predict the time until an event occurs, MTLR divides the time axis into discrete intervals and uses logistic regression to model the probability of an event occurring within each interval. This effectively allows for a non-parametric form of survival modeling that can capture complex and nonlinear relationships between time and event occurrence. MTLR functions by extending logistic regression to multiple correlated tasks, where each task corresponds to the event occurrence within a specific time interval. This multi-task framework enables the sharing of information across different time intervals, thereby providing more robust and informed predictions. Unlike methods like Cox regression, which assume proportional hazards, MTLR does not make specific assumptions about the underlying hazard function. This flexibility allows MTLR to adapt to various data structures and makes it capable of modeling complex time-to-event patterns. In the context of this study, MTLR offers an innovative approach that bridges the gap between traditional survival analysis and binary classification models. Its ability to handle time-to-event data in discrete intervals, coupled with the flexibility to capture nonlinear patterns, makes it a powerful tool for survival analysis. Its use in the given work demonstrates an advanced and nuanced approach to understanding survival dynamics, enriching the analytical toolbox available for time-to-event studies.

Survival methods represent a versatile and robust set of tools for analyzing time-to-event data. From foundational techniques to more sophisticated approaches involving covariates and deep learning integration, survival analysis

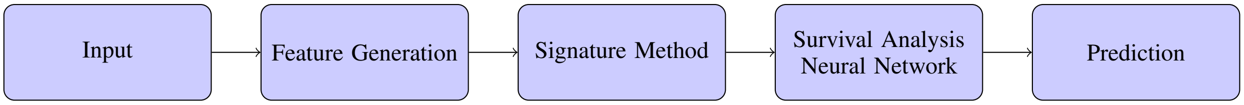


Figure 1: Architecture of survival analysis over signature transformations method showing high-level steps in the method used in this paper (adapted from Kidger et al. [2019])

offers a comprehensive framework for understanding and predicting events over time. Its adaptability and robust handling of censored data make it an invaluable methodology for research and practice across various disciplines. By continually evolving and integrating with contemporary technologies, survival analysis remains at the forefront of statistical modeling and predictive analytics.

2 Methods

The proposed architecture, shown in figure 1 forms a systematic pipeline for predicting survival analysis, integrating various computational techniques. The initial stage of the architecture begins with the raw input data, potentially consisting of patient information or relevant factors. This data is processed through the feature generation block, where essential characteristics are extracted and transformed into a structured representation. Subsequently, the signature method translates this structured data into a signature using iterated integrals, capturing intricate time dependencies within the data.

The signature is then input into a survival analysis neural network, a specially designed block focusing on survival analysis tasks, including predicting time-to-event outcomes. This part of the architecture can be tailored using various configurations and optimization techniques to suit the specific nature of the survival analysis. Finally, the processed information is transformed into the final prediction in the prediction block, providing insightful information about the survival analysis, such as probabilities or expected survival times.

The described architecture emphasizes a synergy of traditional feature engineering, mathematical abstraction through signature methods, and modern deep learning techniques. By integrating these various approaches, the system aims to create a robust and flexible platform capable of handling complex survival analysis tasks.

2.1 Data Preprocessing

The Kaggle sepsis dataset offers a collection of clinical data aimed at bolstering research on early sepsis prediction. This dataset is derived from the MIMIC-III (Medical Information Mart for Intensive Care III) database [Johnson et al., 2016], a large, publicly available dataset developed by the MIT Lab for Computational Physiology. The MIMIC-III dataset itself comprises de-identified health data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

The Kaggle sepsis dataset is a frequently used asset in machine learning research [Devi et al., 2022]. It has been instrumental in the algorithmic evolution related to sepsis, serving both as a training ground and a test bed. Researchers leverage this dataset to calibrate, evaluate, and innovate on a diverse range of algorithms, from traditional statistical models to intricate deep learning constructs.

Within the Kaggle sepsis subset, data spans a broad spectrum, covering vital signs (*e.g.*, heart rate, pulse oximetry), laboratory values (*e.g.*, bicarbonate levels, arterial blood oxygen saturation), and demographic details (*e.g.*, age, gender, ICU admission details). The primary outcome variable, `SepsisLabel`, is tailored to identify septic patients, marking a 1 for timestamps six hours prior to a clinical sepsis prediction. The values for each feature are marked at hourly intervals for each patient, though no value may be available.

The data preparation was produced in several steps. First, the data for all patients was loaded and the number of missing values for each feature calculated. Those columns with greater than 95 percent missing values were noted. Then the mean value of each column, excluding missing values were calculated. Finally, a time point is identified at which all patients will be evaluated. This is described in terms of number of hours after entry into the intensive care unit (ICU) and is referred to as the test point. The time points used in this analysis are $T = \{3, 6, 12, 18, 24\}$.

Then each patient's data was processed individually. First, if a patient were to become positive for sepsis, the time difference between onset and the test point is identified. If the patient were not to become septic, the time difference between the last observation and the test point is used as the censoring time. Then all observations for the patient after the test point are removed. If there are fewer than two observations remaining, the patient is dropped from the dataset.

Then, the data is expanded to add rows for hourly intervals for which no observations are available. Missing data, due to lack of availability in the original dataset and a newly created observation time, are filled using the last observation carried forward (LOCF) method [Shao and Zhong, 2003]. There are risks of introducing bias with LOCF, but the method was chosen here for its commonality. At this point, any missing values, either due to the lack of a value for the patient, or those observation points prior the feature's first observation are filled with the feature mean found earlier.

2.2 Signature Transformation

At this point, an individual patient's representation contains no missing values and is passed to the signature method to calculate the signature transformation. The signature step is performed now, during data preparation, rather than during training for two reasons. First, the resultant signature is a single row of transformed features for each patient. Therefore, in a dataset presented for training, each patient is a single row. In addition, the resultant signature does not change unless the underlying data changes and the computational cost of recomputing the signatures for each training is not considered worth it.

The final observation for each patient then consists of a sepsis label, that is, true or false, the time of onset or censoring, the age of the patient, sex of the patient, which are static during the analysis. To this information is added the signature and the data is saved for training.

2.3 Application of Neural Networks

These datasets are then trained using five different neural survival models, listed above and included in the PyCox package for neural survival model [Kvamme, 2021]. These are, with their data encodings:

1. Cox regression (`coxtime`)
2. Cox proportional hazard (`coxph`)
3. PC-Hazard regression (`pchazard`)
4. partial logistic regression (`logistic`)
5. multi-task logistic regression (MTLR) (`mtlr`)

These trainings were conducted 240 times for each of the five neural models and each of the five-time points for a total of 6000 training sessions. To accommodate such a large number of neural trainings, these were distributed across the Open Science Grid (OSG) [Pordes et al., 2007, Sfiligoi et al., 2009, OSG, 2006] which completed the task in less than 12 hours using CPU-only training.

2.4 Evaluation Methods

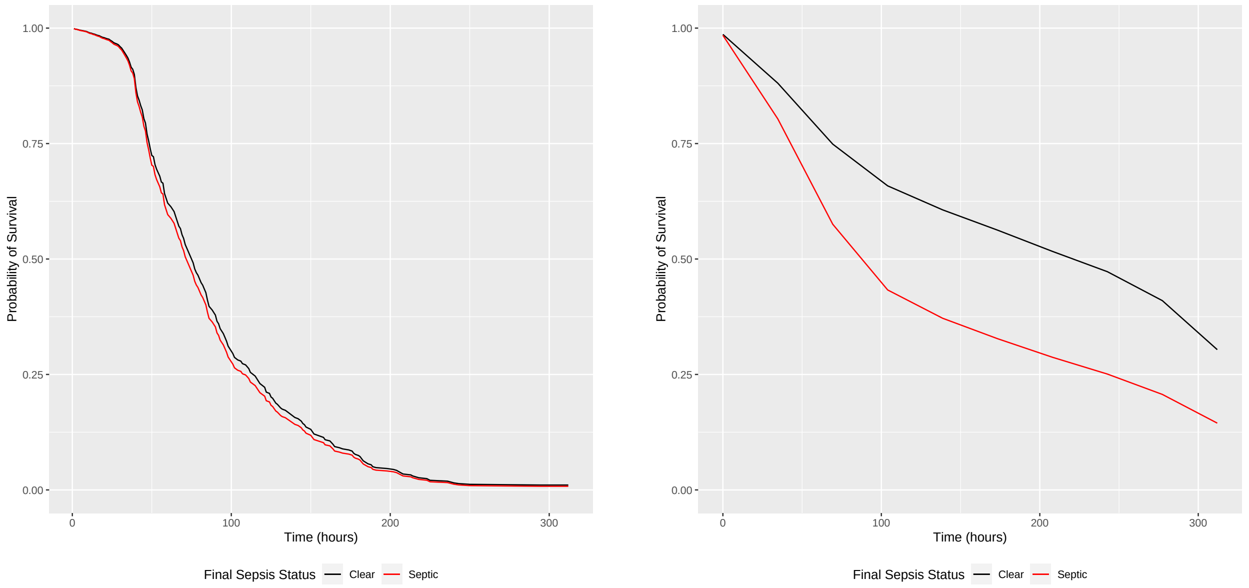
To evaluate each model, we use three different metrics. Firstly, we adopt the Antolini concordance [Antolini et al., 2005]. Unlike traditional measures that offer a single, fixed metric to evaluate model performance, the Antolini index considers the temporal dynamics inherent in survival data. It evaluates how well a model can distinguish between individuals who experience an event (such as death) at a particular time versus those who do not, considering the entire survival curve. The index ranges from 0 to 1, with values closer to 1 indicating better discriminative ability and values near 0.5 signifying no better than random chance.

Secondly, we adopt the integrated inverse probability of censoring weighting (IPCW) Brier score, used to assess the predictive accuracy of survival models, especially in the context of censored observations [Graf et al., 1999, Gerds and Schumacher, 2006]. The IPCW Brier score adjusts for censoring by employing inverse probability of censoring weighting to incorporate censored instances properly. The index ranges from 0 to 1, with values closer to 0 indicating better discriminative ability.

Thirdly, we adopt the integrated IPCW negative binomial log-likelihood as a specialized measure used to assess the fit and predictive performance of survival models [Graf et al., 1999], particularly when the data may contain overdispersion and censoring. The log-likelihood function quantifies how well the model fits the observed data, with a higher log-likelihood indicating a better fit.

3 Results

To achieve clinical relevance, these predictive models must unequivocally delineate the trajectory of patients who are susceptible to developing sepsis from those who are not. Conventional hazard models are typically based on the



(a) Cox proportional hazard model run number 139

(b) MTLR model run number 139

Figure 2: Examples of convergent and divergent predictions

assumption of right-censoring for all observations, which posits that the event of interestin this case, the onset of sepsisis inevitable but may not occur within the observed timeframe. While this assumption is suitable for events that are almost certain to happen, such as mechanical failure or the natural mortality of organisms, it proves to be a flawed premise when applied to the development of sepsis in patients. This is because sepsis is not an inevitable outcome for all individuals, making the assumption of inevitable occurrence problematic for the evaluation of model efficacy.

To address these complexities, we conducted an in-depth examination of survival curves for each outcome group separately. A model that can significantly differentiate the long-term outcomes between these groups would be indicative of its predictive validity. To illustrate this, we provided visual comparisons of two distinct models in figure 2. In figure 2a, we showcase a Cox proportional hazard that fails to demarcate the survival curves of the two groups effectively, thus questioning its applicability in clinical settings. Conversely, figure 2b displays a MTLR model that exhibits a distinct separation between the survival curves, validating its potential for accurately predicting sepsis onset.

These side-by-side comparisons serve not merely as illustrations but as instructive guides that underscore the subtleties and nuanced criteria essential for model selection in predicting sepsis. In a field where the stakes are highoften a matter of life and deaththe imperativeness of choosing a model with robust predictive accuracy cannot be overstated. Therefore, our illustrative examples should prompt further discussions and rigorous assessments for the continual refinement of predictive models in healthcare analytics.

Bridging the illustrative analyses with the statistical evaluations, it becomes evident that the process of model selection is a multifaceted endeavor requiring a balanced interplay between empirical data and theoretical considerations. Utilizing the Antolini concordance as a specific evaluative metric adds a layer of rigor to the appraisal process, enriching our understanding of model performance over varying time intervals. This shift from a more qualitative assessment to a quantifiable metric allows for a more objective comparison, which is vital for the translation of these models into actionable insights for clinical practice. Therefore, our findings as detailed in table 1, coupled with the comparative survival curves, collectively contribute to the establishment of a more comprehensive framework for model selection in the prediction of sepsis. The Cox proportional hazard model yielded a concordance ranging from 0.395 to 0.667 with an average of 0.548, displaying peak performance at the 18.000-hour mark. In contrast, the Cox regression model surpassed Cox proportional hazard in performance with a minimum concordance of 0.421, an average of 0.570, and a maximum of 0.679, exhibiting robustness across time frames, including at the 24 time point.

Among other models, the partial logistic regression model's performance was relatively modest, spanning from 0.376 to 0.587 with an average of 0.484. This result was notably consistent across different intervals. The MTLR model, while showing a wider concordance range between 0.415 and 0.643, reached a high concordance of 0.641 at the 18-hour

Time (hours)	Minimum	Average	Maximum
3	0.468	0.553	0.624
6	0.462	0.553	0.652
12	0.430	0.554	0.652
18	0.395	0.547	0.667
24	0.434	0.534	0.631

(a) Cox proportional hazards model

Time (hours)	Minimum	Average	Maximum
3	0.461	0.566	0.624
6	0.449	0.585	0.654
12	0.448	0.567	0.649
18	0.470	0.578	0.660
24	0.421	0.555	0.679

(b) Cox model

Time (hours)	Minimum	Average	Maximum
3	0.431	0.502	0.575
6	0.386	0.489	0.573
12	0.409	0.479	0.566
18	0.386	0.477	0.587
24	0.376	0.471	0.577

(c) Partial logistic regression model

Time (hours)	Minimum	Average	Maximum
3	0.450	0.504	0.578
6	0.415	0.520	0.594
12	0.435	0.531	0.617
18	0.470	0.551	0.641
24	0.458	0.557	0.643

(d) Multi-task logistic regression (MTLR) model

Time (hours)	Minimum	Average	Maximum
3	0.034	0.055	0.085
6	0.035	0.058	0.084
12	0.021	0.041	0.065
18	0.018	0.036	0.062
24	0.017	0.029	0.065

(e) PC-Hazard model

Table 1: Minimum, average, and maximum Antolini concordance for each model type

time point. However, the PC-Hazard model distinctly lagged behind others, with an exceptionally low concordance, ranging from 0.017 to 0.085, indicating its limited effectiveness.

These results elucidate different survival models' comparative strengths and limitations in predicting sepsis. The Cox regression model is the most promising approach in this context, whereas the PC-Hazard model's underperformance warrants further investigation. The findings highlight the significance of selecting appropriate models and the potential benefit of further optimization, especially at specific time intervals, to enhance prediction accuracy.

The assessment of the integrated IPCW Brier (IBS) further delineates the relative performance of the different modeling approaches in predicting sepsis outcomes, with measurements given in table 2. The Cox proportional hazard model displayed IBS values ranging from 0.102 to 0.215, with an average score of 0.134, and was relatively consistent across varying time intervals. The Cox regression model similarly performed well, with scores ranging from 0.103 to 0.182 and an average of 0.135, exhibiting steady improvement over time with a peak at the 24-hour mark.

The partial logistic regression model demonstrated a wider range of performance, with a minimum IBS of 0.108 and a maximum of 0.225, leading to an average score of 0.147. Although it started at a comparable level to other models, the increase in IBS values at later intervals indicates a potential decrease in predictive accuracy over time. The MTLR model exhibited the highest IBS values among the considered models, ranging from 0.107 to 0.247, with an average of 0.163. This trend culminated at the 24-hour window, suggesting a growing divergence from the actual outcomes as time progressed. Finally, the PC-Hazard model maintained a consistent performance, with values between 0.111 and 0.211, and an average of 0.154.

The evaluation of the integrated IPCW Brier scores highlights the nuanced performance differences among the models. While Cox proportional hazard and CoxTime models provided relatively stable and low IBS across intervals, the MTLR model's increasing IBS signifies potential overfitting or sensitivity to certain variables, warranting further investigation. These insights reinforce the importance of careful model selection and tuning to optimize predictive performance in sepsis prediction.

The integrated IPCW negative binomial log-likelihood (INBLL) scores provide an additional metric for evaluating the performance of the various models, given in table 3. The Cox proportional hazard model revealed INBLL scores ranging

Time (hours)	Minimum	Average	Maximum
3	0.102	0.116	0.149
6	0.106	0.120	0.142
12	0.117	0.132	0.162
18	0.130	0.145	0.174
24	0.138	0.159	0.215

(a) Cox proportional hazards model

Time (hours)	Minimum	Average	Maximum
3	0.103	0.117	0.156
6	0.107	0.122	0.144
12	0.118	0.134	0.152
18	0.128	0.146	0.175
24	0.142	0.159	0.182

(b) Cox model

Time (hours)	Minimum	Average	Maximum
3	0.108	0.123	0.137
6	0.110	0.128	0.146
12	0.127	0.142	0.162
18	0.141	0.158	0.184
24	0.151	0.183	0.225

(c) Partial logistic regression model

Time (hours)	Minimum	Average	Maximum
3	0.107	0.132	0.160
6	0.110	0.141	0.182
12	0.132	0.161	0.189
18	0.143	0.180	0.212
24	0.161	0.202	0.247

(d) Multi-task logistic regression (MTLR) model

Time (hours)	Minimum	Average	Maximum
3	0.111	0.135	0.163
6	0.113	0.141	0.174
12	0.126	0.153	0.183
18	0.138	0.163	0.194
24	0.150	0.176	0.211

(e) PC-Hazard model

Table 2: Minimum, average, and maximum integrated IPCW Brier score for each model type

from 0.329 to 0.849, with an average value of 0.415. The scores consistently increased over time, reaching a peak at the 24-hour interval. The Cox regression model demonstrated a similar pattern, with scores ranging from 0.333 to 0.633 and an average of 0.420.

In contrast, the partial logistic regression model showed INBLL values between 0.357 and 0.699, with an average of 0.464. This progression reflects an upward trend over time, with the highest values observed at later intervals. The MTLR model exhibited the most substantial variation in INBLL scores, from 0.377 to 0.865, with an average of 0.562. This upward trajectory culminated at the 24-hour mark, indicating a potential increase in the divergence from the true outcomes over time. Lastly, the PC-Hazard model maintained scores between 0.455 and 0.775, with an average of 0.577, showing a consistent increase across the time intervals.

Analyzing the INBLL scores uncovers distinct performance patterns among the considered models. While Cox proportional hazard and Cox regression models maintained relatively steady and low INBLL scores, the MTLR and PC-Hazard models exhibited higher values and significantly increased over time. These trends suggest variations in model fit and predictive accuracy, emphasizing the necessity of cautious interpretation and selection depending on the specific context and objectives of the sepsis prediction.

Finally, for each model, we calculated the mean predicted probability of sepsis 96 hours after the time point for both the sepsis-negative (x) and sepsis-positive groups (y). Further, the statistical difference between these means was tested (p-value), and the area under the curve (AUC) was calculated for those predictions. These are summarized in table 4.

For the Cox proportional hazard model, the x ranged around 0.987, and the y was approximately 0.986, with p-values ranging from 0.147 to 0.259. The AUC varied from 0.430 to 0.631, averaging at 0.547. Similarly, the Cox regression model exhibited an x of about 0.989 and y of 0.987, with p-values between 0.117 and 0.222, and an AUC average of 0.567.

The partial logistic regression model's x and y were substantially lower, around 0.005, with p-values ranging from 0.258 to 0.553, and AUC values averaging at 0.528. In contrast, the MTLR model showed a more significant difference between x (approximately 0.200) and y (around 0.111), with p-values as low as 9.12E-07, and an AUC averaging at

Time (hours)	Minimum	Average	Maximum
3	0.329	0.371	0.507
6	0.341	0.380	0.472
12	0.363	0.406	0.514
18	0.393	0.442	0.698
24	0.415	0.475	0.849

(a) Cox proportional hazards model

Time (hours)	Minimum	Average	Maximum
3	0.333	0.378	0.630
6	0.346	0.387	0.473
12	0.365	0.414	0.500
18	0.389	0.443	0.633
24	0.425	0.475	0.597

(b) Cox model

Time (hours)	Minimum	Average	Maximum
3	0.357	0.408	0.457
6	0.370	0.421	0.488
12	0.398	0.454	0.528
18	0.437	0.490	0.580
24	0.468	0.549	0.699

(c) Partial logistic regression model

Time (hours)	Minimum	Average	Maximum
3	0.377	0.459	0.557
6	0.396	0.491	0.630
12	0.457	0.556	0.657
18	0.468	0.617	0.772
24	0.511	0.685	0.865

(d) Multi-task logistic regression (MTLR) model

Time (hours)	Minimum	Average	Maximum
3	0.455	0.547	0.672
6	0.459	0.542	0.646
12	0.469	0.565	0.688
18	0.501	0.592	0.746
24	0.544	0.638	0.775

(e) PC-Hazard model

Table 3: Minimum, average, and maximum integrated IPCW negative binomial log-likelihood for each model type

0.644. Finally, the PC-Hazard model's x and y were around 0.008, with p-values ranging from 0.338 to 0.674, and an average AUC of 0.518. The grand total of the observed values indicated an overall x of 0.438 and y of 0.420, with an average AUC of 0.561.

The diverse patterns across different models indicate varying capabilities in predicting sepsis 96 hours post time point. While Cox proportional hazard and Cox regression maintained closely aligned mean predicted probabilities for both groups, models like MTLR exhibited significant differences, reflected in the low p-values. The range of AUC values across models also highlights the differences in overall prediction performance, emphasizing the need for careful model selection based on the specific requirements of the sepsis prediction task.

4 Discussion

The evaluation of various predictive models demonstrates nuanced performance in terms of Antolini concordance, integrated IPCW Brier scores, and integrated IPCW negative binomial log-likelihood scores. The Cox regression model consistently displayed superior performance across different time intervals, with partial logistic regression and MTLR models following closely. The distinct performance characteristics of each model might be attributed to underlying algorithmic differences, leading to unique predictive behaviors.

The Antolini concordance values indicate a performance spread across models, ranging from 0.017 for PC-Hazard to 0.679 for Cox regression. This suggests differential abilities to accurately order the risk of events, with the Cox regression model displaying a pronounced ability to rank the risk of sepsis in patients. The differences in these values highlight the importance of selecting the appropriate model for specific clinical scenarios.

The Brier scores further highlight the calibration of the models, with Cox proportional hazard and Cox regression models demonstrating better calibration compared to logistic and PC-Hazard models. This reflects how well the probabilities are aligned with the observed outcomes, which is a vital aspect of predictive accuracy, especially in the context of medical diagnoses where precision is paramount.

Time (hours)	x	y	p-values	Minimum AUC	Average AUC	Maximum AUC
3	0.982	0.980	0.147	0.470	0.557	0.615
6	0.985	0.984	0.182	0.447	0.548	0.603
12	0.988	0.987	0.153	0.430	0.549	0.620
18	0.990	0.989	0.176	0.433	0.545	0.631
24	0.990	0.990	0.259	0.446	0.538	0.610

(a) Cox proportional hazards model

Time (hours)	x	y	p-values	Minimum AUC	Average AUC	Maximum AUC
3	0.984	0.982	0.122	0.479	0.569	0.625
6	0.988	0.986	0.136	0.458	0.577	0.635
12	0.990	0.989	0.159	0.435	0.557	0.632
18	0.991	0.990	0.117	0.502	0.572	0.654
24	0.991	0.991	0.222	0.464	0.558	0.639

(b) Cox model

Time (hours)	x	y	p-values	Minimum AUC	Average AUC	Maximum AUC
3	0.006	0.005	0.258	0.437	0.539	0.606
6	0.005	0.005	0.318	0.427	0.539	0.628
12	0.005	0.005	0.553	0.424	0.515	0.625
18	0.005	0.005	0.544	0.425	0.521	0.626
24	0.004	0.004	0.488	0.449	0.524	0.599

(c) Partial logistic regression model

Time (hours)	x	y	p-values	Minimum AUC	Average AUC	Maximum AUC
3	0.150	0.100	0.011	0.517	0.610	0.664
6	0.190	0.114	0.005	0.530	0.640	0.684
12	0.218	0.119	0.003	0.548	0.653	0.699
18	0.233	0.116	0.000	0.605	0.662	0.703
24	0.208	0.105	0.000	0.584	0.653	0.715

(d) Multi-task logistic regression (MTLR) model

Time (hours)	x	y	p-values	Minimum AUC	Average AUC	Maximum AUC
3	0.010	0.010	0.338	0.445	0.536	0.608
6	0.009	0.009	0.436	0.427	0.535	0.605
12	0.009	0.009	0.647	0.406	0.505	0.609
18	0.008	0.009	0.674	0.427	0.505	0.595
24	0.007	0.007	0.631	0.429	0.509	0.591

(e) PC-Hazard model

Table 4: Minimum, average, and maximum AUC for 96-hour predictions for each model type

In terms of the integrated IPCW negative binomial log-likelihood, the MTLR model exhibited the maximum score of 0.865, whereas the Cox proportional hazard model had the minimum score of 0.329. This signals a difference in goodness-of-fit with potential implications for the selection of models for specific clinical settings. Understanding these differences can guide clinicians in selecting the model that best suits their patients' needs.

Analyzing the mean predicted probabilities of sepsis for the negative and positive groups, the Cox proportional hazard and Cox regression models demonstrated closer mean predictions between the groups, but with higher AUC values. This reflects better discrimination, while logistic and PC-Hazard models presented lower AUC values, suggesting less effectiveness in distinguishing between the two groups.

The p-values provided indicate the statistical significance of the mean differences between the two groups, with the MTLR model showing highly significant differences, as seen by low p-values, while the PC-Hazard model exhibited less significance. This highlights the ability of some models to differentiate between sepsis and non-sepsis groups in a statistically robust manner, which is key in medical research.

The evaluation of these models in predicting sepsis has profound implications for clinical decision-making processes. The judicious selection of a model, substantiated by robust performance metrics, has the potential to revolutionize the timely and precise diagnosis of sepsis. This, in turn, could serve as a critical factor in enhancing patient outcomes and reducing mortality rates. The seamless integration of these predictive models into existing healthcare infrastructures highlights their pragmatic applicability, offering an invaluable tool for clinicians in real-world scenarios.

Interestingly, the performance of the predictive models varied across different time intervals, which further necessitates a nuanced, context-dependent approach to their implementation. This variation accentuates the importance for healthcare professionals to be well-versed in the intricacies and limitations of each model. Given that patient conditions are inherently dynamic, fluctuating due to a myriad of variables such as treatment responses or the progression of underlying conditions, the 'one-size-fits-all' strategy is inadequate. Hence, healthcare providers must exercise discernment in applying these models, tailoring their use to fit the specific needs and circumstances of individual patient cases. This calls for a more comprehensive understanding of how these models operate within the ever-changing landscape of clinical care.

Finally, we recognize important to recognize that the study's conclusions are based on a specific dataset and conditions, with generalizability to other populations and settings needing further exploration. Future research may focus on integrating diverse clinical variables, refining model calibration, and evaluating these models in prospective clinical trials to provide robust, actionable tools for healthcare providers.

The architecture presented in this study represents an exciting convergence of traditional feature engineering, mathematical abstraction, and contemporary neural network techniques. The integration of signature methods to transform features allows the model to capture complex temporal dynamics that might otherwise be missed by conventional methods. This innovative blend not only offers a fresh perspective on survival analysis but also extends its predictive capabilities. By recognizing patterns over time, this approach fosters more accurate risk assessments and personalized predictions, vital in various applications such as medical prognosis and financial forecasting.

One of the architecture's salient features is its modular design, ensuring that it can be tailored and adapted to different domains and datasets. Each component, from feature generation to neural network structure, can be optimized to suit the specific problem, endowing the architecture with a broad range of applications. This flexibility makes it a versatile tool that can be fine-tuned to meet the unique requirements of researchers and practitioners across different fields.

In the realm of computational efficiency, the signature method exhibits noteworthy advantages that hold practical implications for real-world applications. Unlike more computationally-intensive deep learning algorithms that necessitate specialized hardware like graphics processing units (GPUs) for effective deployment, the signature method's computational demands are surprisingly moderate. This is primarily because the algorithm focuses on transformational mathematics that generate meaningful but relatively sparse feature sets, as opposed to methods that require extensive backpropagation through potentially vast neural networks. Consequently, the computational load is lessened to a degree where a standard laptop central processing unit (CPU) is entirely adequate for both training and inference tasks. This operational efficiency not only reduces associated computational costs but also broadens the range of devices on which the model can be run, thereby democratizing access to advanced predictive analytics. This inherently lean computational profile makes the signature method a compelling choice for settings where computational resources are limited, yet high-accuracy predictions are paramount.

The integration of signature methods within the survival analysis pipeline is particularly novel. By transforming structured data into signatures using iterated integrals, the architecture introduces a level of mathematical sophistication often lacking in purely data-driven models. This infusion of mathematical rigor enhances both the performance and interpretability of the model, shedding light on the underlying mechanisms that guide its predictions.

Looking ahead, the proposed architecture lays a promising foundation for more intelligent and nuanced survival analysis. Future research could explore enhancements such as advanced neural network configurations, alternative signature methods, or the incorporation of additional data sources. Comparative studies with existing models will further delineate the specific advantages and potential limitations of this approach. Beyond its technical attributes, the architecture's potential impact on healthcare, finance, and other sectors is profound. Its capacity for more precise and individualized predictions supports improved decision-making, risk management, and resource allocation.

In sum, the architecture offers an inspiring blend of innovation, adaptability, and far-reaching potential. By leveraging the strengths of diverse computational techniques and introducing a new method for survival analysis, it paves the way for continued exploration and broad application. The methodology described in this paper sets a precedent for interdisciplinary collaboration, uniting mathematical elegance with machine learning's predictive prowess. Its implications resonate across a myriad of fields, positioning it as a valuable asset in the ever-evolving landscape of data science.

Acknowledgements

This research was done using services provided by the OSG Consortium [Pordes et al., 2007, Sfiligoi et al., 2009, OSG, 2006], which is supported by the National Science Foundation awards #2030508 and #1836650.

References

Anne Hunt. Sepsis: an overview of the signs, symptoms, diagnosis, treatment and pathophysiology. *Emergency Nurse*, 30(5), 2022.

Patrick Kidger and Terry Lyons. Signatory: differentiable computations of the signature and logsignature transforms, on both cpu and gpu. *arXiv preprint arXiv:2001.00706*, 2020.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of machine learning research*, 20(129):1–30, 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):1–12, 2018.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime data analysis*, 27:710–736, 2021.

Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–1186, 1998.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in neural information processing systems*, volume 24, 2011.

Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, 2018.

Terry J Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310, 1998.

Terry Lyons and Zhongmin Qian. Flow equations on spaces of rough paths. *journal of functional analysis*, 149(1): 135–159, 1997.

Weixin Yang, Lianwen Jin, Hao Ni, and Terry Lyons. Rotation-free online handwritten character recognition using dyadic path signature features, hanging normalization, and deep neural network. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4083–4088. IEEE, 2016.

Anh Tong, Thanh Nguyen-Tang, Toan Tran, and Jaesik Choi. Learning fractional white noises in neural stochastic differential equations. *Advances in Neural Information Processing Systems*, 35:37660–37675, 2022.

Christian Bayer, Peter K Friz, and Nikolas Tapia. Stability of deep neural networks via discrete rough paths. *SIAM Journal on Mathematics of Data Science*, 5(1):50–76, 2023.

Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *arXiv preprint arXiv:1810.10971*, 2018.

Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*, 2016.

Patrick Kidger, Patric Bonnier, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep signature transforms. *Advances in Neural Information Processing Systems*, 32, 2019.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

V Aruna Devi, Sakthi Jaya Sundar Rajasekar, and Varalakshmi Perumal. Sepsense: A novel sepsis detection system using machine learning techniques. *Healthcare 4.0: Health Informatics and Precision Data Management*, page 131, 2022.

Jun Shao and Bob Zhong. Last observation carry-forward and last observation analysis. *Statistics in medicine*, 22(15): 2429–2441, 2003.

Håvard Kvamme. Survival analysis with pytorch, 2021. URL `https://github.com/havakv/pycox`.

Ruth Pordes, Don Petravick, Bill Kramer, Doug Olson, Miron Livny, Alain Roy, Paul Avery, Kent Blackburn, Torre Wenaus, Frank Würthwein, Ian Foster, Rob Gardner, Mike Wilde, Alan Blatecky, John McGee, and Rob Quick. The open science grid. In *J. Phys. Conf. Ser.*, volume 78 of *78*, page 012057, 2007. doi:10.1088/1742-6596/78/1/012057.

Igor Sfiligoi, Daniel C Bradley, Burt Holzman, Parag Mhashilkar, Sanjay Padhi, and Frank Wurthwein. The pilot way to grid resources using glideinwms. In *2009 WRI World Congress on Computer Science and Information Engineering*, volume 2 of *2*, pages 428–432, 2009. doi:10.1109/CSIE.2009.950.

OSG. OSPool, 2006. URL `https://osg-htc.org/services/open_science_pool.html`.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, 24(24):3927–3944, 2005.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Thomas A Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.