An evidential time-to-event prediction model based on Gaussian random fuzzy numbers

L. Huang, Yucheng Xing, Thierry Denoeux, Mengling Feng

An evidential time-to-event prediction model based on Gaussian random fuzzy numbers

Ling Huang[1][0000−0003−1562−1371], Yucheng Xing[1], Thierry Denœux[3,4][0000−0002−0660−5436], and Mengling Feng[1,2][0000−0002−5338−6248]

[1] Saw Swee Hock School of Public Health, National University of Singapore, Singapore
huang.l@nus.edu.sg
[2] Institute of Data Science, National University of Singapore, Singapore
[3] Université de technologie de Compiègne, CNRS, Heudiasyc, France
[4] Institut Universitaire de France, France

Abstract. We introduce an evidential model for time-to-event prediction with censored data. In this model, uncertainty on event time is quantified by Gaussian random fuzzy numbers, a newly introduced family of random fuzzy subsets of the real line with associated belief functions, generalizing both Gaussian random variables and Gaussian possibility distributions. Our approach makes minimal assumptions about the underlying time-to-event distribution. The model is fit by minimizing a generalized negative log-likelihood function that accounts for both normal and censored data. Comparative experiments on two real-world datasets demonstrate the very good performance of our model as compared to the state-of-the-art.

Keywords: Survival analysis · belief functions · Dempster-Shafer theory · random fuzzy sets · uncertainty · machine learning.

1 Introduction

Time-to-event analysis, also known as survival analysis, focuses on analyzing the time it takes for an event of interest to occur, such as time to death or machine failure. The main challenge of time-to-event prediction is that the observed outcomes are typically censored, meaning that the exact event time is unknown for some data due to early-end experiments or a lack of follow-up, making the estimation problem challenging. Conventional statistical machine learning techniques focus on the estimation of the hazard function, mathematically defined as the ratio of the probability density to the time-to-event function, representing the conditional probability density that a single nonrepeatable event will occur in a particular time interval, given that the item did not experience the event before that time.

The Cox proportional hazards model (Cox model) [1], proposed by Cox in 1972, offers a straightforward approach to handling censoring by assuming proportional hazards across covariates while leaving the baseline hazard function

unspecified. Faraggi and Simon [8] introduced an extension of the Cox model by replacing its linear predictor with a one-hidden layer multilayer perceptron (MLP). Recent advancements of the Cox model with deep neural networks, e.g., DeepSurve [10] and Cox-CC [11], show promising performance. However, the proportional hazards assumption may not hold in complex scenarios, potentially leading to biased estimates and inaccurate predictions. To address this limitation, Kvamme et al. [11] proposed a time-dependent Cox model to account for time-varying covariates. Furthermore, the Cox-based model estimates the baseline hazard function solely based on observed event times, which can introduce extra biases or information loss when data is limited. Random Survival Forests (RSF) [9], a non-parametric model that builds upon the random forest algorithm and ensemble learning, shows advantages where traditional parametric or semi-parametric methods may not be suitable or when the underlying survival distribution is complex. In addition to estimating the time-to-event distribution, recent deep-learning advanced approaches also focus on improving prediction performance with new optimization strategies. For instance, DeepHit [12], a probability mass function-based discrete-time model, shows promising concordance index results with a loss function designed to improve event time ranking while disregarding the calibration of the predictions.

In this paper, we propose an evidence-based time-to-event prediction model that does not rely on specific forms of data distribution assumptions. Instead, we calculate the evidence of a time interval directly under the framework of belief functions [2,13] and random fuzzy sets [3,7]. The proposed approach modifies the ENNreg model introduced in [4,6] to account for censored data. Prediction uncertainty is quantified using Gaussian random fuzzy numbers (GRFNs) [7], a newly introduced family of random fuzzy subsets of the real line. In addition to providing the most plausible event time, our model outputs two additional quantities: standard deviation and precision measuring, respectively, aleatory and epistemic prediction uncertainties. The model is fitted by minimizing a generalized negative log-likelihood loss function.

The rest of this paper is organized as follows. Background notions are first recalled in Section 2. The proposed model is then introduced in Section 3, and experimental results are reported in Section 4. Finally, Section 5 concludes the paper.

2 Background

The theory of epistemic random fuzzy sets (RFSs) was introduced in [3,7] as an extension of Dempster-Shafer theory allowing us to represent both partially reliable and vague evidence. In short, an RFS is a mapping from a probability space to the fuzzy powerset of another space, verifying some measurability property. The reader is referred to the cited references for a general exposition of this theory. Hereafter, we briefly recall the notions of Gaussian and lognormal random fuzzy numbers in Sections 2.1 and 2.2, respectively.

2.1 Gaussian random fuzzy numbers

A Gaussian fuzzy number (GFN) is a fuzzy subset of the real line with membership function $x \mapsto \exp(-0.5h(x-m)^2)$, where $m \in \mathbb{R}$ and $h \geq 0$ are the mode and precision parameters. A Gaussian random fuzzy number (GRFN) $\widetilde{T}$ is an RFS defined as a GFN whose mode M is a Gaussian random variable with mean μ and variance σ^2 [7]. It is then defined by three parameters μ, σ^2 and h and we write $\widetilde{T} \sim \widetilde{N}(\mu, \sigma^2, h)$. The family of GRFNs is closed under the product-intersection rule, a combination operator generalizing Dempter's rule [7]. A GRFN defines a belief function of the real line. Formulas for the degrees of belief and plausibility of any real interval are given in [7].

2.2 Lognormal random fuzzy numbers

A GRFN is a model of uncertainty about a variable taking values in the whole real line. In contrast, in time-to-event analysis, the response variable is positive. Uncertainty about such a variable is better represented by a lognormal random fuzzy number as introduced in [5].

In general, let ψ be a one-to-one mapping from $\mathbb{R}$ to a subset $\Lambda \subseteq \mathbb{R}$. Its extension $\widetilde{\psi}$ maps each fuzzy subset $\widetilde{F}$ of $\mathbb{R}$ to a fuzzy subset $\widetilde{\psi}(\widetilde{F})$ of Λ with membership function $\lambda \mapsto \widetilde{F}[\psi^{-1}(\lambda)]$. Let $[0,1]^{\mathbb{R}}$ denote the set of all fuzzy subsets of $\mathbb{R}$, and $\widetilde{Y} : \Omega \to [0,1]^{\mathbb{R}}$ be a RFS. By composing $\widetilde{\psi}$ with $\widetilde{Y}$, we obtain a new RFS $\widetilde{\psi} \circ \widetilde{Y} : \Omega \to [0,1]^{\Lambda}$. For any event $C \subseteq \Lambda$, we have

$$Bel_{\widetilde{\psi}\circ\widetilde{Y}}(C) = Bel_{\widetilde{Y}}(\psi^{-1}(C)) \quad \text{and} \quad Pl_{\widetilde{\psi}\circ\widetilde{X}}(C) = Pl_{\widetilde{Y}}(\psi^{-1}(C)). \tag{1}$$

Taking $\widetilde{Y} \sim \widetilde{N}(\mu, \sigma^2, h)$, $\Lambda = [0, +\infty)$ and $\psi = \exp$, we obtain a *lognormal random fuzzy number* $\widetilde{T} : \widetilde{\exp} \circ \widetilde{Y}$ denoted by $\widetilde{T} \sim T\widetilde{N}(\mu, \sigma^2, h, \log)$. We can remark that $\widetilde{\log} \circ \widetilde{Y} \sim \widetilde{N}(\mu, \sigma^2, h)$. Degrees of belief $Bel_{\widetilde{T}}(I)$ and $Pl_{\widetilde{T}}(I)$ for any interval $I \subseteq [0, +\infty)$ can easily be computed from (1) and formulas given in [7] for GRFNs.

3 Model

Our approach is based on the ENNreg model introduced in [6]. This model will be recalled in Section 3.1. The loss function adapted to censored data will then be described in Section 3.2.

3.1 Evidential time-to-event prediction network

In time-to-event analysis, the response of an event time is always positive, while a GRFN is a model about a variable taking values in the whole real line. Following the idea of *Lognormal random fuzzy numbers* as we introduced in Section 2.2, we construct a transformed GRFN-based evidential neural network to map predictions into the positive real timeline with $Y = \log(T)$, where T is the time

to event. Here the network is composed of three layers: the distance layer, the evidence mapping layer, and the fusion layer. The distance layer computes the distances between the input vector x and each prototype p_k with a positive scale parameter γ_k: $s_k(x) = \exp(-\gamma_k^2 \|x - p_k\|^2)$. For each prototype p_k, the evidence mapping computes a GRFN $\widetilde{N}(\mu_k(x), \sigma_k^2, s_k(x)h_k)$, where σ_k^2 and h_k are variance and precision parameters, and $\mu_k(x)$ is given by $\mu_k(x) = \beta_k^T x + \beta_{k0}$, where β_k is a p-dimensional vector of coefficients and β_{k0} is a scalar parameter. The evidence fusion layer combines evidence from the K prototypes using the unnormalized product-intersection combination rule $\boxplus$ [7] and outputs a final GRFN $\widetilde{Y}(x) \sim \widetilde{N}(\mu(x), \sigma^2(x), h(x))$ given by

$$\mu(x) = \frac{\sum_{k=1}^{K} s_k(x) h_k \mu_k(x)}{\sum_{k=1}^{K} s_k(x) h_k}, \quad \sigma^2(x) = \frac{\sum_{k=1}^{K} s_k^2(x) h_k^2 \sigma_k^2}{(\sum_{k=1}^{K} s_k(x) h_k)^2},$$

and $h(x) = \sum_{k=1}^{K} s_k(x) h_k$. Output $\mu(x)$ denotes the most plausible time-to-event prediction, $\sigma^2(x)$ denotes the variance around $\mu(x)$ (aleatory uncertainty), and $h(x)$ denotes the precision of the prediction (epistemic uncertainty). Uncertainty about T is then described by the lognormal RFN $\widetilde{T} \sim T\widetilde{N}(\mu(x), \sigma^2(x), h(x), \log)$.

3.2 Loss function

To optimize the proposed framework, we use negative generalized log-likelihood loss defined in [6], and adapt it to account for both uncensored and censored data. If the data is not censored, the continuous event time $\widetilde{Y}$ is always observed with finite precision. Therefore, instead of observing an exact value, we actually observe an interval $[y]_\epsilon = [y - \epsilon, y + \epsilon]$ centered at y. Our prediction evidence can, therefore, be characterized by either the degree of belief $Bel([y]_\epsilon)$ or the plausibility $Pl([y]_\epsilon)$. Conversely, if the data is censored, the event time $\widetilde{Y}$ will be observed in interval $[y, \infty)$. Our prediction evidence can now be represented as the degree of belief $Bel([y, \infty))$ or plausibility functions $Pl([y, \infty))$ in the time interval $[y, \infty)$. We can optimize the time-to-event function based either on L_{Bel} or L_{Pl}. While none of these two functions adequately measures the quality of the imprecise predictions as mentioned in [6]. Let $\widetilde{Y}$ be the output GRFN, $y = \log(t)$ the observation, and D a binary censoring variable such that $D = 1$ if $Y = y$, and $D = 0$ if it is only known that $Y \geq y$. We, therefore, consider the following weighted sum of L_{Bel} or L_{Pl}

$$\mathcal{L}_{\lambda,\epsilon}(\widetilde{Y}, y, D) = \lambda \overline{\mathcal{L}}(\widetilde{Y}, y, D) + (1 - \lambda) \underline{\mathcal{L}}(\widetilde{Y}, y, D),$$

with

$$\overline{\mathcal{L}}(\widetilde{Y}, y, D) = -D \ln Bel_{\widetilde{Y}}([y - \epsilon, y + \epsilon]) - (1 - D) \ln Bel_{\widetilde{Y}}([y, \infty)),$$

and

$$\underline{\mathcal{L}}(\widetilde{Y}, y, D) = -D \ln Pl_{\widetilde{Y}}([y - \epsilon, y + \epsilon]) - (1 - D) \ln Pl_{\widetilde{Y}}([y, \infty)),$$

where λ is the hyperparameter that controls the cautiousness of the prediction (the smaller, the more cautious). We set $\lambda = 0.1$ to enable the model to focus more on plausibility optimization. The hyperparameter ϵ was set to 10^{-6} to present an infinitesimal time interval.

The network is trained by minimizing the regularized average loss

$$C^{(R)}_{\lambda,\epsilon,\xi,\rho}(\Psi) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\lambda,\epsilon}(\widetilde{Y}(x_i;\Psi), y_i, D_i) + \frac{\xi}{K}\sum_{k=1}^{K}h_k + \frac{\rho}{K}\sum_{k=1}^{K}\gamma_k^2,$$

where Ψ is the vector of all parameters (prototypes are included as well) in the network, $\widetilde{Y}(x_i;\Psi)$ is the network output for input x_i, and ξ, ρ are two regularization parameters. The first regularizer term has the effect of reducing the number of prototypes used for the prediction (e.g., setting $h_k = 0$ to discarding prototype k), while the second regularizer term shrinks the solution towards a linear model (e.g., setting $\gamma_k = 0$ for all k yields a linear model) In [6], ξ and ρ are tuned by cross-validation. In the experiments reported in Section 4, we kept them fixed at $\xi = \rho = 0.1$ for simplicity.

4 Experimental results

We will now show some qualitative results of our method applied to a simulated dataset with various data censoring scenarios (Section 4.1), and compare its performance to state-of-the-art time-to-event prediction methods on two real-world datasets (Section 4.2).

4.1 Illustrative example on simulated dataset

We first consider artificial data with the following distribution: the input X has a uniform distribution in the interval $[-2, 2]$, and the response is

$$T = \exp\left[1.5X + 2\cos(3X)^3 + \frac{X+5}{3\sqrt{5}}V\right], \tag{2}$$

where $V \sim N(0,1)$ is a standard normal random variable. To simulate data censoring scenarios, two elements were incorporated: the event censoring state indicator D and a random censoring value C. The event indicator D has a Bernoulli distribution, denoted as $D \sim B(p)$, where $1-p$ represents the censoring rate (set as 0.1 and 0.7). For events flagged with a censoring indicator $D = 1$, a negative value C is added to T to emulate right censoring. Here, the value C follows a uniform distribution, with $C \sim U(-1, 0)$ and $C \sim U(-2, 0)$ representing different degrees of censoring severity. Learning and validation sets of size $n = 4000$ and $n = 1000$ were generated.

The model was initialized with $K = 40$ prototypes. The targets $y = \log(t)$, the network outputs $\mu(x)$, along with belief prediction intervals (BPIs) at levels $\alpha \in \{0.5, 0.9, 0.99\}$ are shown in Fig. 1. BPIs, as defined in [6], are intervals

centered at $\mu(x)$ with the degree of belief α to contain the true value of the response variable. When only 10% of the data are censored, our model predicts a time-to-event function (red line) that closely aligns with the ground truth function (blue broken line), and the predicted BPIs effectively encompass the majority of data points, as shown in Fig. 1a. With 70% of the data censored (Fig. 1b), the predicted time-to-event function becomes smoother with fewer details and exhibits an upward bias relative to the true sample distribution, as expected. Nevertheless, the BPIs still effectively encompass the majority of data points, though they are wider. When the censoring interval increases, for example, from $[0, 1]$ (Fig. 1a) to $[0, 2]$ (Fig.1c), our model continues to perform well with even wider BPIs.

We can conclude that for different data censoring scenarios, the predicted time-to-event functions closely model the actual regression function, even when the data is highly censored. Furthermore, the BPIs effectively encompass the majority of data points. These observations illustrate the robustness of our approach to varying data censoring conditions.

4.2 Comparative results on real-world datasets

We further evaluated our approach using two real-world time-to-event datasets provided by [10]. These are the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset, comprising 1904 samples with a censoring rate of 42%, and the Rotterdam Tumor Bank and German Breast Cancer Study Group (GBSG) dataset, containing 2232 samples with a censoring rate of 43%. Following [11], we used the concordance index (C_{idx}) to evaluate the prediction performance, as well as the integrated Brier score (IBS) and the integrated binomial log-likelihood (IBLL) to evaluate the calibration of the estimates. We used five-fold cross-validation and repeated it five times. We compared our methods to the baseline Cox method, RSF [9], Deepsurv [10], Cox-CC [11], Cox-Time [11] and DeepHit [12]. Hyperparameter values for the compared methods are given in the documentation of the Pycox package[5].

Results for the two datasets are reported in Tables 1 and 2. We can remark that all the compared methods are based on specific assumptions, and it is not surprising that some of them perform quite well for specific data distribution. The Cox model performs rather poorly, which was expected as it is based on very restrictive model assumptions. The methods that assume proportional hazards without linearity assumption, i.e., Cox-CC and DeepSurv perform worse, in general, than the less restrictive methods, namely, RFS and Deephit.

Our ENNreg method achieved the best performance according to the C_{idx} and IBLL criteria, and the second best. Notably, our proposal outperforms the continuous Cox-time model by a large amount and performs slightly better than the discrete DeepHit model. This result is interesting considering that we did not develop a time-dependent prediction model like Cox-time, nor did we use concordance for hyperparameter tuning as in Deephit. As we can see from the IBS and

[5] https://github.com/havakv/pycox

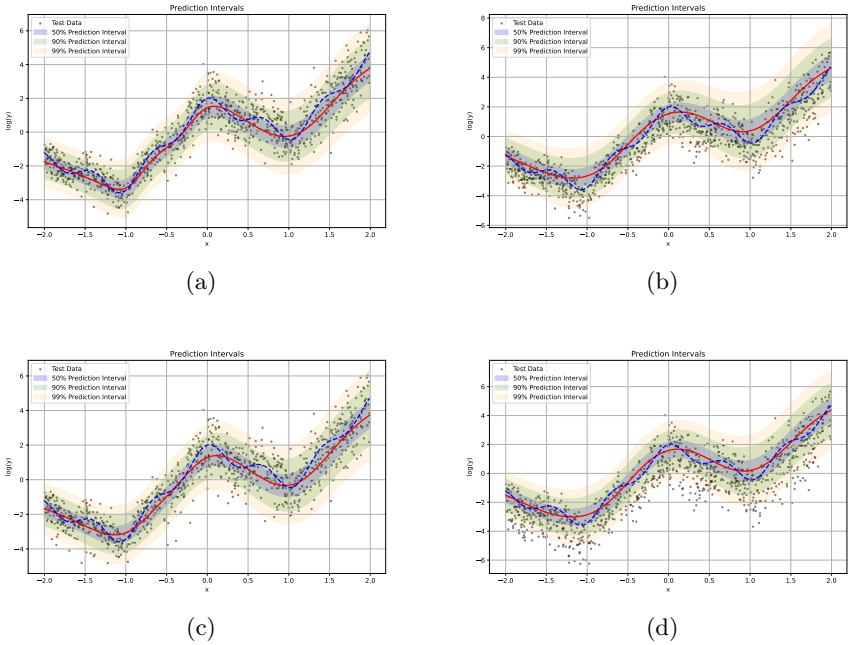

Fig. 1: Simulated data, actual regression function (blue broken lines), and predictions obtained from the trained model. Predicted values $\log(y)$ are depicted by red solid lines, while belief prediction intervals (BPIs) at levels $\alpha \in \{0.5, 0.9, 0.99\}$ are represented by shaded areas in blue, green, and orange. The first and second rows are data with censoring intervals $[-1, 0]$ and $[-2, 0]$, respectively. The first and second columns are data with 10% and 70% censoring rates, respectively.

IBLL results, the promising concordance performance of Deephit comes at the cost of poorly calibrated survival estimates. In contrast, our proposal exhibits good calibration properties, with statistically significant differences observed in calibrated survival estimates for both datasets. We can, therefore, conclude that our evidence-based time-to-event prediction model, based on minimal assumptions, demonstrates greater flexibility and robustness compared to state-of-the-art models that rely on restrictive hypotheses such as the proportional hazard assumption.

5 Conclusion

In time-to-event analysis, some proportion of the data is usually censored. In this paper, we have adapted the ENNreg model introduced in [6] to account for censored data, and applied it to time-to-event prediction. The model is trained

Table 1: Means and standard errors of C_{idx}, IBS and IBLL scores on the Metabric database for our method (ENNreg) and alternative methods. The best and second best results are, resp., in bold and underlined.

Methods	$C_{idx}\uparrow$	IBS$\downarrow$	IBLL $\downarrow$
Cox	$0.633\pm9.3\times10^{-3}$	$0.164\pm3.3\times10^{-3}$	$0.497\pm1.1\times10^{-2}$
RFS	$0.644\pm1.2\times10^{-3}$	$0.173\pm0.9\times10^{-3}$	$0.510\pm2.0\times10^{-3}$
Deepsurv	$0.646\pm7.4\times10^{-3}$	$\mathbf{0.162}\pm3.6\times10^{-3}$	$\underline{0.493}\pm1.2\times10^{-2}$
Cox-cc	$0.641\pm2.1\times10^{-3}$	$\underline{0.163}\pm3.3\times10^{-3}$	$\mathbf{0.490}\pm8.6\times10^{-3}$
Cox-time	$\underline{0.663}\pm1.0\times10^{-2}$	$0.164\pm4.6\times10^{-3}$	$0.488\pm1.1\times10^{-2}$
DeepHit	$\mathbf{0.672}\pm1.0\times10^{-2}$	$0.173\pm2.6\times10^{-3}$	$0.516\pm6.5\times10^{-3}$
ENNreg	$\mathbf{0.672}\pm9.4\times10^{-3}$	$\underline{0.163}\pm2.1\times10^{-3}$	$\mathbf{0.490}\pm5.0\times10^{-3}$

Table 2: Means and standard errors of C_{idx}, IBS and IBLL scores on the GBSG database for our method (ENNreg) and alternative methods. The best and second best results are, resp., in bold and underlined.

Methods	$C_{idx}\uparrow$	IBS$\downarrow$	IBLL $\downarrow$
Cox	$0.669\pm2.5\times10^{-3}$	$\mathbf{0.174}\pm3.3\times10^{-3}$	$\underline{0.519}\pm1.7\times10^{-3}$
RFS	$0.655\pm0.3\times10^{-3}$	$0.190\pm0.5\times10^{-3}$	$0.539\pm1.0\times10^{-3}$
Deepsurv	$0.666\pm8.4\times10^{-3}$	$0.180\pm1.9\times10^{-3}$	$0.531\pm5.1\times10^{-3}$
Cox-cc	$0.672\pm3.3\times10^{-3}$	$\mathbf{0.174}\pm0.5\times10^{-3}$	$0.529\pm3.4\times10^{-3}$
Cox-time	$\underline{0.678}\pm4.7\times10^{-3}$	$\underline{0.177}\pm1.5\times10^{-3}$	$0.523\pm3.7\times10^{-3}$
DeepHit	$\underline{0.678}\pm4.5\times10^{-3}$	$0.195\pm1.0\times10^{-3}$	$0.565\pm2.6\times10^{-3}$
ENNreg	$\mathbf{0.681}\pm2.2\times10^{-3}$	$\mathbf{0.174}\pm1.1\times10^{-3}$	$\mathbf{0.518}\pm2.8\times10^{-3}$

using the logarithm of the response variable T as the target variable and outputs a GRFN. Prediction uncertainty is, thus, quantified by a lognormal random fuzzy number, from which degrees of belief and plausibility of various events can be straightforwardly computed. In this paper, we have focused on prediction accuracy and calibration (assessed using standard performance criteria) and showed the good performance of our model on two datasets as compared to the state-of-the-art. In the future, we will further explore the advantages of uncertainty quantification in time-to-event tasks using belief functions, e.g., studying the standard deviation and precision of the prediction. We will also extend the comparison with state-of-the-art to a wider range of clinical medical datasets for different time-to-event tasks. Moreover, the study of mixtures of GFRNs to fit applications, e.g., finance analysis, should also be interesting.

6 Acknowledgment

This research is supported by A*STAR, CISCO Systems (USA) Pte. Ltd, and National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002) and the National Research Foundation Singapore under AI Singapore Programme (Award AISG-GC-2019-001-2B).

References

1. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)
2. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics **38**, 325–339 (1967)
3. Denœux, T.: Belief functions induced by random fuzzy sets: A general framework for representing uncertain and fuzzy evidence. Fuzzy Sets and Systems **424**, 63–91 (2021)
4. Denœux, T.: An evidential neural network model for regression based on random fuzzy numbers. In: Le Hégarat-Mascle, S., Bloch, I., Aldea, E. (eds.) Belief Functions: Theory and Applications. pp. 57–66. Springer International Publishing, Cham (2022)
5. Denœux, T.: Parametric families of continuous belief functions based on generalized gaussian random fuzzy numbers. Fuzzy Sets and Systems **471**, 108679 (2023)
6. Denœux, T.: Quantifying prediction uncertainty in regression using random fuzzy sets: the ennreg model. IEEE Transactions on Fuzzy Systems (2023)
7. Denœux, T.: Reasoning with fuzzy and uncertain evidence using epistemic random fuzzy sets: General framework and practical models. Fuzzy Sets and Systems **453**, 1–36 (2023)
8. Faraggi, D., Simon, R.: A neural network model for survival data. Statistics in medicine **14**(1), 73–82 (1995)
9. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. Annals of Applied Statistics **2**(2), 841–860 (2008)
10. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology **18**(1), 1–12 (2018)
11. Kvamme, H., Borgan, Ø., Scheel, I.: Time-to-event prediction with neural networks and cox regression. Journal of machine learning research **20**(129), 1–30 (2019)
12. Lee, C., Zame, W., Yoon, J., Van Der Schaar, M.: Deephit: A deep learning approach to survival analysis with competing risks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
13. Shafer, G.: A mathematical theory of evidence, vol. 42. Princeton university press (1976)