Atelier Data Science

M2 MIASHS

Table des matières

Novembre 2025

Informations générales

Ce projet s'inscrit dans la prolongation du module d'Analyse de Survie Pénalisée du Master 2 MIASHS. Il implique une démarche d'investigation personnnelle autour de la problématique de la **prédiction conforme en analyse de survie**.

Le jeu de données provient du projet **SigBERT** [1], dont le dépôt GitHub est disponible à l'adresse suivante : `https://github.com/MINCHELLA-Paul/SigBERT`. Les étudiants disposent pour cet atelier du sous-répertoire *Atelier_SigBERT* au lien `https://github.com/MINCHELLA-Paul/Master-MIASHS/tree/main/Atelier_SigBERT`. Les datasets zipés prêt à utiliser sont `df_study_selected.zip` et `df_study_selected_L36_w6.zip` et rassemblent l'ensemble des observations prétraitées et prêtes à être analysées dans un cadre de survie.

1 Structure de projet proposée

Le travail se déroule en plusieurs étapes logiques, permettant d'aboutir à une analyse complète de la performance et de la fiabilité du modèle de survie.

1.1 Présentation du Dataset

Chaque ligne du dataset choisi correspond à un individu anonymisé, caractérisé par trois composantes principales. La première est un identifiant unique `ID` permettant de distinguer les patients. La deuxième regroupe un ensemble de coefficients de signatures, notés $\mathbb{S} = \left(S^{(1)}, \ldots, S^{(p)}, S^{(1,1)}, \ldots, S^{(p,p)}\right)$, qui sont extraits à partir des embeddings textuels cliniques générés par SigBERT. Ces signatures ont le bon goût de capter l'évolution temporelle du suivi des patients à travers les sentence embeddings représentatif des textes cliniques ; ordonnés dans le temps. La troisième composante rassemble les variables de survie : un indicateur binaire d'événement `event` (valant 1 en cas de décès, 0 sinon) et une durée de suivi `time` exprimée en unités temporelles appropriées.

Les signatures $\mathbb{S}$ seront les covariables utilisées dans le modèle de Cox pénalisé. Le score de risque linéaire de ce modèle s'écrit $\eta = \beta \cdot \mathbb{S}$, où β représente le vecteur des coefficients de régression à estimer. Cette formulation permet de quantifier l'influence de chaque composante des signatures sur le risque instantané de survenue de l'événement d'intérêt.

1.2 Définition de la tâche

L'objectif de ce travail est double :

1. Reproduire l'ajustement d'un modèle de survie sur le dataset fourni. Plusieurs approches sont envisageables : un modèle de Cox pénalisé classique, un réseau de neurones estimant directement le score de risque, une forêt aléatoire adaptée aux données de survie, ou tout autre type de modèle de survie ne reposant pas nécessairement sur l'hypothèse de proportionnalité des risques de Cox. L'estimation du score de risque individuel $\hat{\eta}$ pour chaque patient constitue le premier livrable, accompagnée d'une évaluation de la capacité discriminante du modèle à l'aide de métriques standards telles que le C-index ou l'AUC dépendant du temps (td-AUC), mais aussi une étude de sa calibration.
2. Proposer et discuter une approche de **prédiction conforme** adaptée à ce contexte de survie. La prédiction conforme vise à quantifier l'incertitude des prédictions de

manière rigoureuse, en fournissant des garanties de couverture statistique. Plusieurs stratégies peuvent être explorées :

- → Produire des intervalles de confiance autour du score de risque estimé $\hat{\eta}$, permettant ainsi de mesurer la variabilité de cette prédiction ponctuelle.
- → Quantifier l'incertitude sur une probabilité de survie à un horizon temporel fixé, notée $\mathbb{P}(T_i > t^\star \mid \mathbb{S}_i)$, ce qui offre une interprétation plus directe pour les cliniciens.
- → Construire des ensembles de prédiction conformes fournissant une borne inférieure pour le temps de survie individuel T_i. Concrètement, pour un nouveau patient caractérisé par ses signatures $\mathbb{S}_i$, l'objectif est de déterminer un seuil temporel $\underline{t}_i(\alpha)$ tel que $\mathbb{P}(T_i \geq \underline{t}_i(\alpha)) \geq 1 - \alpha$ avec une garantie de couverture marginale contrôlée au niveau $1 - \alpha$. Cette borne minimale présente un intérêt clinique direct : elle permet d'affirmer avec un niveau de confiance spécifié qu'un patient survivra au moins jusqu'à un horizon temporel donné. La construction de tels intervalles repose sur l'utilisation d'un score de non-conformité adapté au contexte censuré, par exemple basé sur les résidus de Cox ou sur des pseudo-observations dérivées des courbes de survie prédites. L'évaluation de cette approche porte à la fois sur la couverture empirique effective (vérification que la proportion de patients survivant au-delà de leur borne atteint bien le niveau nominal) et sur l'efficacité des prédictions (mesurée par la valeur moyenne ou médiane des bornes inférieures obtenues).

Le choix de la cible à laquelle appliquer la prédiction conforme est volontairement laissé ouvert. Il est attendu que la pertinence de l'approche proposée soit **solidement argumenté** en tenant compte de la nature du modèle sélectionné, de la faisabilité technique de la méthode conforme, et de l'interprétabilité clinique des résultats obtenus. Cette réflexion méthodologique constitue une partie essentielle de l'évaluation.

1.3 Modélisation et algorithmes

La modélisation repose sur plusieurs étapes méthodologiques complémentaires. Premièrement, le choix du modèle (Cox classique, Cox pénalisé, Cox Neural Network) estimant directement le score de risque, ou toute autre approche de modélisation de survie jugée pertinente par l'équipe. Le choix du modèle doit être justifié en fonction de la dimensionnalité des signatures, du nombre d'événements observés, et des hypothèses sous-jacentes.

Ensuite, pour tout modèle comportant un paramètre de régularisation (comme dans le cas du Cox pénalisé), une validation croisée doit être mise en œuvre pour sélectionner de manière optimale ce paramètre, en maximisant par exemple le C-index sur les plis de validation, ou en minimisant le Brier Score à un instant choisi.

Une fois le modèle calibré, on dispose de chaque $\hat{\eta}_i = \hat{\beta} \cdot \mathbb{S}_i$ pour tout patient i.

Enfin, l'application d'une méthode de prédiction conforme permet de quantifier l'incertitude des prédictions individuelles avec des garanties statistiques rigoureuses. Plusieurs stratégies peuvent être adoptées :

1. **Conformalized Survival Analysis (CSA)** pour produire des intervalles de prédiction sur le temps de survie T_i de chaque individu, avec une garantie de couverture marginale au niveau $1 - \alpha$ souhaité,
2. **Conformalized Survival Distributions (CSD)** pour construire des bandes de confiance autour des fonctions de survie individuelles $S_i(t)$, permettant ainsi de quantifier l'incertitude sur l'ensemble de la trajectoire temporelle,

3. ou encore une adaptation méthodologique propre à l'équipe étudiante, par exemple une approche basée sur la calibration des probabilités prédites, une méthode par simulation Monte Carlo, ou une combinaison originale de techniques existantes. Toute proposition alternative doit être rigoureusement justifiée et accompagnée d'une validation empirique. *Un ensemble de méthodes possibles est proposé dans la table 1.*

L'objectif n'est pas uniquement de coder une procédure, mais de **relier la prédiction conforme au concept d'incertitude en médecine** : quelles garanties offre-t-elle, et que permet concrètement de proposer une couverture conforme dans un contexte de survie ?

1.4 Évaluation et discussion

Les performances du modèle seront évaluées via des métriques standards : C-index pour la discrimination globale, td-AUC pour la discrimination temporelle ; score de Brier et courbes de calibration pour la justesse des probabilités prédites. Il est nécessaire de **discuter de l'intérêt de ces métriques**.

L'apport de la prédiction conforme devra être analysé en vérifiant la validité empirique des intervalles (taux de couverture effectif vs. nominal $1 - \alpha$) et en examinant l'impact sur les performances discriminantes. La discussion portera sur l'interprétation clinique des ensembles de prédiction, les limites méthodologiques (hypothèses d'échangeabilité, gestion de la censure, coût computationnel), et des pistes d'amélioration possibles telles que la stratification du risque ou la calibration post-hoc.

2 Livrables attendus

Le travail donnera lieu à deux productions principales. Le premier livrable consiste en un **notebook Jupyter applicatif**, rédigé en Python ou R, documentant l'ensemble de la démarche méthodologique. Ce notebook devra inclure le prétraitement des données, l'ajustement du modèle de survie avec justification des choix effectués, l'implémentation complète de la méthode de prédiction conforme retenue, l'évaluation des performances à l'aide des métriques appropriées, ainsi que des visualisations claires et informatives (courbes de Kaplan-Meier, distributions des scores de risque, intervalles conformes, courbes de calibration). Le code devra être propre, commenté et reproductible, avec une gestion explicite des dépendances logicielles.

Le second livrable est une **présentation orale** (slides en PowerPoint ou PDF) de 15 minutes (± 2 minutes), suivie de 5 minutes de questions. Cette présentation, appuyée par des supports visuels synthétiques (slides), devra couvrir de manière équilibrée la problématique et les enjeux de la prédiction conforme en survie, la présentation du dataset et du modèle retenu, les résultats obtenus avec des visualisations percutantes, ainsi qu'une discussion critique des performances et des limites de l'approche.

L'analyse doit être rédigée sous forme de note synthétique (2–3 pages) accompagnée d'un notebook reproductible.

3 Références proposées

Publication	Résumé de la tâche et de la méthode	Ref.
Conformalized Survival Analysis	Introduction du cadre fondateur de la prédiction conforme en survie, visant à produire des bornes inférieures prédictives pour le temps de survie sous censure à droite. La méthode est sans hypothèse de distribution et fournit une couverture marginale exacte pour tout modèle de survie (Cox, RSF, etc.).	[2]
Conformalized Survival Distributions	Propose un post-traitement conforme pour les distributions de survie (*Conformal Survival Distributions*), améliorant la calibration empirique des probabilités de survie tout en maintenant la discrimination des modèles existants.	[3]
Survival Conformal Prediction Under Random Censoring	Étend la CSA au cadre de la censure aléatoire. L'article construit des intervalles prédictifs bilatéraux pour le temps de survie, en adaptant les statistiques de non-conformité pour gérer les observations partiellement censurées.	[4]
Two-Sided Conformalized Survival Analysis	Présente la version bilatérale de la Conformalized Survival Analysis, permettant d'obtenir à la fois des bornes inférieure et supérieure sur les temps de survie prédits.	[5]
Doubly Robust Conformalized Survival Analysis	Introduit une approche doublement robuste combinant imputation et pondération inverse pour la Conformalized Survival Analysis, améliorant la validité sous censure informative.	[6]
Conformalized Survival Analysis for General Right-Censoring	Généralise la CSA au cadre de censure à droite non uniforme. Propose des scores de non-conformité individuels pour ajuster la couverture empirique.	[7]
Conformal Survival Bands for Risk Screening	Construit des bandes conformes pour les courbes de survie individuelles, afin d'évaluer visuellement l'incertitude associée aux prédictions, notamment pour les applications de dépistage du risque.	[8]

TABLE 1 – Publications majeures sur la prédiction conforme appliquée à l'analyse de survie.

Conclusion

Cet atelier vise à confronter les étudiants à une problématique moderne : *comment intégrer une incertitude mathématiquement garantie dans un modèle de survie issu de l'apprentissage automatique*, dont le questionnement aujourd'hui est primordial pour une étude clinique complète.

Références

[1] Paul Minchella, Loic Verlingue, Stéphane Chrétien, Rémi Vaucher, and Guillaume Metzler. Sigbert : Combining narrative medical reports and rough path signature theory for survival prediction in oncology. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track. ECML PKDD 2025*, volume part IX of *Lecture Notes in Computer Science*, Porto, Portugal, September 2025. Springer. URL https://ecmlpkdd-storage.s3.eu-central-1.amazonaws.com/preprints/2025/ads/preprint_ecml_pkdd_2025_ads_778.pdf.

[2] Emmanuel Candès, Jing Lei, and Zhen Ren. Conformalized survival analysis. *arXiv preprint arXiv :2103.09763*, 2021. URL https://arxiv.org/abs/2103.09763.

[3] Hanwen Qi, Haoxuan Yu, and Russell Greiner. Conformalized survival distributions : A generic post-process to increase calibration. *arXiv preprint arXiv :2405.07374*, 2024. URL https://arxiv.org/html/2405.07374v1.

[4] Shuo Yi, Rui Song, and Lili Wang. Survival conformal prediction under random censoring. *Stat*, 14(2) :e70052, 2025. doi : 10.1002/sta4.70052. URL https://onlinelibrary.wiley.com/doi/10.1002/sta4.70052.

[5] Shuqing Xu, Zhen Ren, and Emmanuel Candès. Two-sided conformalized survival analysis. *arXiv preprint arXiv :2403.00841*, 2024. URL https://arxiv.org/abs/2403.00841.

[6] Ying Zhou, Yifan Zhang, Yunzhe Li, and Kun He. Doubly robust conformalized survival analysis with right-censored data. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.00597.

[7] Jiacheng Wang, Zhen Ren, and Jing Lei. Conformalized survival analysis for general right-censoring. *OpenReview preprint*, 2025. URL https://openreview.net/forum?id=CSA-RC-2025.

[8] Wenqi Zhao, Mengdi Liu, and Junjie Li. Conformal survival bands for risk screening under right-censoring. *arXiv preprint arXiv :2504.00456*, 2025. URL https://arxiv.org/abs/2504.00456.