Avis de Soutenance

Monsieur Imad EL BADISY

RECHERCHES BIOMEDICALES Santé publique

Soutiendra publiquement ses travaux de thèse intitulés

Données manquantes et apprentissage automatique

dirigés par Monsieur Roch GIORGI

Soutenance prévue le ***mardi 09 décembre 2025*** à 9h00
Lieu : SESSTIM SITE SANTÉ TIMONE (QuanTIM - SanteRCom) Faculté de Médecine 27 Bd Jean Moulin 13385 Marseille Cedex 5.
Salle : de visio-conférence

Composition du jury proposé

M. Roch GIORGI	Aix-Marseille Université (Faculté de Médecine) / Assistance Publique - Hôpitaux de Marseille	Directeur de thèse
M. Yohann FOUCHER	Université de Poitiers	Rapporteur
M. Raphaël PORCHER	Université Paris Cité	Rapporteur
M. Xavier PAOLETTI	University of Versailles St Quentin / Paris Saclay & Institut Curie	Président
Mme Karen LEFFONDRE	Université de Bordeaux, ISPED	Examinatrice

Mots-clés : analyse de survie,imputation des données manquantes,machine learning,,

Résumé :

Les données manquantes de covariables sont omniprésentes en analyse de survie et peuvent biaiser les estimations, réduire l'efficacité statistique et altérer l'interprétation clinique. La méthode d'imputation multiple (MI) constitue la stratégie de référence pour traiter la non-réponse, mais ses performances sont généralement évaluées à l'aide de critères restreints, tels que le biais ou la couverture des coefficients de régression. Cette perspective centrée sur les paramètres devient néanmoins insuffisante à mesure que l'analyse de survie adopte des modèles plus flexibles et des méthodes d'apprentissage automatique capables de capturer des effets non linéaires et dépendants du temps. Ces modèles ne produisent pas de résumés simples de type coefficients, et leur comportement peut être fortement influencé par la méthode d'imputation utilisée. Cette thèse apporte trois contributions principales. Premièrement, elle montre que la performance de l'imputation est intrinsèquement multidimensionnelle : la précision de reconstruction des données, la validité inférentielle et la performance prédictive conduisent souvent à des classements divergents des méthodes. Un cadre d'évaluation multimetrique est donc indispensable. Deuxièmement, à travers des simulations et des études cliniques de cas, elle montre que si les méthodes d'imputation classiques fonctionnent correctement sous l'hypothèse de risques proportionnels et d'effets log-linéaires, elles échouent dans des contextes plus réalistes avec effets

non linéaires ou dépendants du temps. Dans de tels cas, les méthodes d'imputation basées sur l'apprentissage automatique préservent mieux les structures de risque, améliorent la couverture et maintiennent la performance prédictive. Troisièmement, la thèse introduit un nouveau cadre fondé sur la distorsion, qui évalue l'imputation en quantifiant les changements dans les prédictions de survie, les structures d'effet des covariables (via des méthodes d'explicabilité) et les métriques de performance. Cela reformule l'imputation comme un problème de préservation du comportement du modèle plutôt que de récupération de paramètres. Dans l'ensemble, cette thèse vise à contribuer aux fondements méthodologiques de l'analyse de survie avec données incomplètes en intégrant modélisation flexible, évaluation multimetrique et analyse du comportement des modèles imputés, offrant ainsi une base pour des analyses plus fiables et plus transparentes dans des contextes biomédicaux complexes.

LE DOYEN

Georges LEONETTI