



**Prédiction des maladies cardiovasculaires sur 10 ans à l'aide de modèles
d'apprentissage automatique adaptés à l'analyse de survie : application
au score de Framingham simplifié (FRS-SP)**

Mémoire présenté

dans le cadre du programme de maîtrise en informatique
en vue de l'obtention du grade de maître ès sciences (M.Sc.)

PAR

© **AMY LEYE**

Mai 2026

Composition du jury :

Mouhamed Tarik Moutacalli, président du jury, UQAR

Mehdi Adda, directeur de recherche, UQAR

Abdenour Bouzouane, membre interne (programme par extension), UQAC

Dépôt initial le 27 février 2026

Dépôt final le 21 mai 2026

UNIVERSITÉ DU QUÉBEC À RIMOUSKI
Service de la bibliothèque

Avertissement

La diffusion de ce mémoire ou de cette thèse se fait dans le respect des droits de son auteur, qui a signé le formulaire « *Autorisation de reproduire et de diffuser un rapport, un mémoire ou une thèse* ». En signant ce formulaire, l'auteur concède à l'Université du Québec à Rimouski une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de son travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, l'auteur autorise l'Université du Québec à Rimouski à reproduire, diffuser, prêter, distribuer ou vendre des copies de son travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de la part de l'auteur à ses droits moraux ni à ses droits de propriété intellectuelle. Sauf entente contraire, l'auteur conserve la liberté de diffuser et de commercialiser ou non ce travail dont il possède un exemplaire.

REMERCIEMENTS

La rigueur scientifique et les exigences d'un travail de recherche sont souvent au-delà des seules capacités de l'étudiant. Je ne saurais présenter ce mémoire sans rendre gloire à Dieu pour sa bonté de m'avoir donné la force, le courage et la santé durant l'élaboration de ce mémoire de fin d'études.

D'abord, je tiens à exprimer ma profonde gratitude à mon directeur de recherche, le Professeur **Mehdi Adda**, pour son soutien exceptionnel tout au long de ce parcours. Au-delà de son encadrement académique et de son accompagnement moral, sa rigueur scientifique a profondément façonné ma manière de penser et affûté mon esprit critique. Sa patience, sa bienveillance et la confiance qu'il a placée en moi m'ont permis de faire mes premiers pas dans le monde de la recherche universitaire avec assurance. Merci, Professeur, d'avoir cru en moi, d'avoir nourri mes ambitions et guidé mes efforts tout au long de ce processus.

Je tiens également à exprimer ma profonde reconnaissance à mes collaborateurs de recherche, **Xavier Trudel** et **Mahée Gilbert-Ouimet**, pour leur précieux soutien, leurs conseils avisés et leur disponibilité tout au long de ce travail. Je remercie aussi l'ensemble des membres de l'équipe de recherche de l'axe **SPPOS du CHUL de Saint-Sacrement**, plus particulièrement les membres des équipes de **Xavier Trudel : Alexandre et Flore**, dont la collaboration, les échanges stimulants et le soutien constant m'ont permis d'enrichir ma réflexion, de persévérer malgré les obstacles et de mener à bien ce projet de recherche.

Je souhaite aussi à remercier l'ensemble des professeurs du Département de Mathématiques et d'Informatique de l'Université du Québec à Rimouski pour leur dévouement et la qualité de leur enseignement. Chacun de vous a contribué, à sa manière, à façonner la personne et le chercheur que je deviens. Je suis fière de vous, et fière de mon Université. J'exprime aussi ma gratitude aux membres du jury qui ont accepté d'évaluer mon travail avec attention et bienveillance.

Ces années n'auraient toutefois pas eu la même saveur sans le soutien indéfectible de mes proches. Je pense tout particulièrement à ma mère, **Bintou LEYE**, à mon père, **Khalifa**

Ababacar, et à ma grande sœur, **Fatou Bintou**, qui ont toujours été présents, malgré la distance, pour m'écouter, m'encourager et apaiser mes moments de doute.

Enfin, je souhaite remercier du fond du cœur tous mes amis, du Canada, de la France et du Sénégal, avec qui j'ai partagé des moments de joie, de réflexions et parfois d'incertitude autour de nos projets respectifs. Votre amitié, votre écoute et votre soutien ont largement contribué à rendre cette aventure humaine et académique inoubliable.

RÉSUMÉ

Les maladies cardiovasculaires (MCV) restent la principale cause de mortalité mondiale. Plusieurs outils de prédiction du risque cardiovasculaire, dont le score de Framingham (FRS) ont été développés. Ainsi, une version simplifiée du FRS, le score de Framingham simplifié (FRS-SP) reposant sur un ensemble parcimonieux de prédicteurs aisément mesurables, facilite son application en pratique clinique. Bien que le score de risque simplifié de Framingham (FRS-SP) soit utilisé en soins primaires, il ne permet souvent pas de capturer les interactions non linéaires entre les facteurs de risque. Les modèles de survie basés sur l'apprentissage automatique (AA) offrent une alternative prometteuse, mais peu d'études ont comparé la performance des modèles d'AA de survie en ayant recours aux mêmes prédicteurs, dérivés d'outils existants. Avec un cadre rigoureux et reproductible, nous évaluons et comparons la calibration et la discrimination de trois modèles de survie (Random Survival Forests - RSF, Gradient Boosting Survival Analysis - GBSA et DeepSurv) en utilisant les prédicteurs de FRS-SP. Cette étude a analysé 6 883 participants (49,5 % de femmes, âge moyen 47,0 ans) issus de la cohorte prospective (The PROspective Quebec (PROQ) Study on Work and Health), à l'aide des prédicteurs FRS-SP mesurés au départ (1999-2001). Les trois modèles ont été entraînés à l'aide d'une validation croisée avec une optimisation rigoureuse des hyperparamètres (Optuna) et une gestion du déséquilibre des classes. Les performances ont été évaluées à l'aide du C-index, du score de Brier et des graphiques de calibration, mais aussi avec des analyses stratifiées par sexe et par catégories de risque. Les modèles RSF et GBSA ont surpassé DeepSurv grâce à l'ajustement des hyperparamètres pour compenser le déséquilibre des données et ont obtenu des C-index de 0,73 et des scores de Brier de 0,0779 et de 0,0785, respectivement, démontrant ainsi une bonne calibration. A l'inverse, DeepSurv a surestimé le risque (C-index = 0,67, calibration moyenne = 1,71). L'analyse stratifiée selon le sexe a révélé un C-index de 0,72 chez les femmes malgré des taux d'événements plus faibles. Cependant, le modèle DeepSurv a considérablement surestimé le risque. Globalement, les modèles d'ensemble (RSF et GBSA)

ont démontré des performances prédictives supérieures à celles observées dans l'étude originale de FRS-SP.

Mots clés : [Maladies cardiovasculaires, Prédiction du risque, Analyse de survie, Apprentissage automatique de survie, Forêts aléatoires de survie, Apprentissage profond de survie, Survie par gradient boosting]

ABSTRACT

Cardiovascular disease (CVD) remains the leading cause of death worldwide. Several cardiovascular risk prediction tools, including the Framingham Risk Score (FRS), were developed. A simplified version of the FRS, the Framingham Risk Score Simplified (FRS-SP), based on a limited set of easily measurable predictors, facilitates its application in clinical practice. While the simplified Framingham risk score (FRS-SP) is used in primary care, it often fails to capture the nonlinear interactions between risk factors. Survival models based on machine learning (ML) offer a promising alternative, but their comparative performance using standard clinical variables remains poorly studied. Using a rigorous reproducible framework, we evaluate and compare the calibration and discrimination of three ML survival models (Random Survival Forests - RSF, Gradient Boosting Survival Analysis - GBSA, and DeepSurv) against standard FRS-SP predictors. This study analyzed 6,883 participants (49.5% women, average age 47.0) from the prospective PROQ cohort, using FRS-SP predictors measured at baseline (1999–2001). Three survival ML models (RSF, GBSA, and DeepSurv) were trained using 5-fold cross-validation with rigorous hyperparameter optimization (Optuna) and class imbalance handling. Performance was evaluated via C-index, Brier score, and calibration plots, stratified by sex and by risk categories. RSF and GBSA outperformed Deep Learning, benefiting from hyperparameter tuning for data imbalance. RSF and GBSA achieved C-indices of 0.73 and Brier scores of 0.0779 and 0.0785, respectively, showing excellent calibration. Conversely, DeepSurv overestimated risk (C-index = 0.67, mean calibration = 1.71). Sex-stratified analysis revealed robust performance in women (C-index = 0.72) despite lower event rates. However, the DeepSurv model overestimated the risk considerably. Overall, ensemble models (RSF and GBSA) demonstrated superior predictive performance compared to those reported in the original FRS-SP study.

Keywords: [Cardiovascular disease; Risk Prediction; Time-to-Event Analysis; Survival Machine Learning; Random survival forests; Deep learning survival; Gradient boosting survival]

TABLE DES MATIÈRES

REMERCIEMENTS.....	viii
RÉSUMÉ	x
ABSTRACT.....	xiii
TABLE DES MATIÈRES	xv
LISTE DES TABLEAUX	xviii
LISTE DES FIGURES	xx
LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES.....	xxii
INTRODUCTION GÉNÉRALE	1
1.1 CONTEXTE ET PROBLEMATIQUE.....	1
1.2 OBJECTIFS	3
1.3 METHODOLOGIE	4
1.3.1 Devis.....	4
1.3.2 Population de l'étude.....	4
1.3.3 Variables à l'étude.....	5
1.3.4 Analyses	7
1.4 CONTRIBUTIONS	14
1.5 STRUCTURE DU MÉMOIRE	15
CHAPITRE 1 : REVUE DE LA LITTERATURE.....	17
1.6 CONCEPTS FONDAMENTAUX.....	18
1.6.1 Définition des maladies cardiovasculaires	18
1.6.2 Épidémiologie des maladies cardiovasculaires	18
1.6.3 Risque cardiovasculaire.....	19
1.6.4 Facteurs de risque cardiovasculaires	19
1.6.5 Prédiction du risque cardiovasculaire.....	20
1.7 REVUE DE LA LITTERATURE.....	28
1.7.1 Stratégie de recherche documentaire	28
1.7.2 Études antérieures.....	31

CHAPITRE 2 : SURVIVAL MACHINE LEARNING FOR 10-YEAR CARDIOVASCULAR RISK PREDICTION USING FRAMINGHAM SCORE SIMPLIFIED : A CANADIAN COHORT STUDY	46
2.1 RÉSUMÉ EN FRANÇAIS DE L'ARTICLE	46
CONCLUSION GÉNÉRALE	66
2.2 SYNTHÈSE DES RESULTATS	66
2.3 LIMITATIONS	67
2.4 OBJECTIFS FUTURS	68
RÉFÉRENCES BIBLIOGRAPHIQUES	70

LISTE DES TABLEAUX

Tableau 1. <i>Stratégie de recherche documentaire (PubMed, 2015–2025)</i>	29
Tableau 2. <i>Stratégie de recherche documentaire (PubMed, 2015–2025) (Suite)</i>	30
Tableau 3. <i>Études sur des cohortes de petite taille dans la prédiction du risque cardiovasculaire</i>	34
Tableau 4. <i>Études sur des cohortes de taille moyenne dans la prédiction du risque cardiovasculaire</i>	35
Tableau 5. <i>Études sur des cohortes de grandes taille dans la prédiction du risque cardiovasculaire</i>	37
Tableau 6. <i>Études utilisant les maladies cardiovasculaires globales (MCV) comme issue de survie</i>	37
Tableau 7. <i>Études utilisant des Accidents Vasculaire Cérébraux (AVC) comme issue de survie</i>	39
Tableau 8. <i>Études utilisant des événements cardiovasculaires majeurs indésirables (MACE) comme issue de survie</i>	40
Tableau 9. <i>Études a événements cardiovasculaires athérosclérotiques (ASCVD comme issue de survie</i>	40
Tableau 10. <i>Études utilisant l'Insuffisance cardiaque (IC) comme issue de survie</i>	41
Tableau 11. <i>Études utilisant les maladies coronariennes comme issue de survie</i>	41
Tableau 12. <i>Synthèse des approches méthodologiques et performances des modèles (Quelques études utilisant des techniques de sélection et ingénierie de variables)</i>	42
Tableau 13. <i>Synthèse des approches méthodologiques et performances des modèles (Quelques études utilisant l'optimisation des hyperparamètres + la validation croisée)</i>	43
Tableau 14. <i>Synthèse des approches méthodologiques et performances des modèles (Quelques études sans l'optimisation des hyperparamètres et sans sélection et ingénierie des variables)</i>	44

LISTE DES FIGURES

Figure 1. <i>Méthodologie proposée pour une prédiction du risque MCV</i>	8
Figure 2. <i>Distribution des classes</i>	9
Figure 3. <i>Distribution des classes avant et après le sous-échantillonnage</i>	11
Figure 4. <i>Prédiction du risque</i>	20
Figure 5. <i>Types d'Apprentissage Automatique</i>	23

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

AA	Apprentissage Automatique
RSF	Random Survival Forests
GBSA	Gradient Boosting Survival Analysis
FRS	Score de Risque de Framingham
FRS-SP	Score de Risque de Framingham Simplifié
CPH	Modèle de Cox à risques proportionnels (Cox Proportional Hazards)
PCE	Calculateur d'Équations de Cohorte Regroupées
MCV	Maladies Cardiovasculaires
PROQ	Étude prospective québécoise sur le travail et la santé
IC	Insuffisance cardiaque
AVC	Accident Vasculaire Cérébral
MACE	Événements Cardiovasculaires Majeurs
AIT	Accident Ischémique Transitoire
ASCVD	Événements cardiovasculaires athérosclérotiques
CIM	Classification internationale des maladies
CCP	Classification canadienne des procédures
CCI	Classification canadienne des interventions

INTRODUCTION GÉNÉRALE

1.1 CONTEXTE ET PROBLEMATIQUE

Les maladies cardiovasculaires (MCV) représentent la principale cause de mortalité et de morbidité à l'échelle mondiale (Damen et al., 2016). Elles demeurent une préoccupation majeure en raison de leur taux de mortalité élevé, avec près de 17,9 millions de décès chaque année, soit 31 % de l'ensemble des décès dans le monde (OMS, 2022).

Au Canada, les MCV arrivent au second rang des causes les plus fréquentes de décès en 2008, ce qui représente plus de 64000 Canadiens (Statistique Canada, 2011). C'est dans cette même logique qu'en 2009, 1.6 millions de canadiens ont déclaré d'avoir des troubles cardiaques (Boursalie et al., 2015). Elles représentent également un lourd fardeau sanitaire et économique: près du tiers des décès sont attribuables à une crise cardiaque ou un accident vasculaire cérébral (AVC), et environ 16 % des hospitalisations sont dues à ces maladies avec un coût annuel dépassant 22 milliards de dollars (Smith, 2009). Dans ce contexte, une évaluation du risque des MCV est nécessaire pour une meilleure orientation des mesures préventives. Cela souligne l'importance de disposer de modèles de prédiction fiables et applicables en contexte clinique, permettant d'identifier les individus les plus susceptibles de développer des MCV (Lloyd-Jones, 2010). Les lignes directrices canadiennes (C-CHANGE 2022) en prévention primaire en cardiologie recommandent l'utilisation de modèles de prédiction afin d'estimer le risque cardiovasculaire pour guider la prévention et pour uniformiser la prise en charge clinique (Jain et al., 2023). Parmi les modèles, le score de risque de Framingham (FRS) demeure l'un des scores de risque le plus recommandé par la Société Canadienne de Cardiologie (Pearson et al., 2021) et le plus largement utilisé en Amérique du Nord (Artigao-Rodenas et al., 2013 ; Sud et al., 2022). Le FRS estime le risque de MCV sur 10 ans à partir d'un ensemble de facteurs de risque cardiovasculaires reconnus : l'âge, le sexe, la pression artérielle systolique, le cholestérol total, le cholestérol HDL, le tabagisme et le diabète (Cybulska & Kłosiewicz-Latoszek, 2019). Afin d'en faciliter l'usage clinique, dans un contexte de soins primaires, une version simplifiée, le score de

Framingham-Simplifié (FRS-SP) a été développée. Elle substitue l'indice de masse corporelle (IMC) aux mesures de cholestérol, permettant une estimation rapide du risque sans analyses de laboratoire. Selon l'étude de validation originale, la performance prédictive du modèle FRS-SP est comparable à celle du modèle classique incluant les lipides (D'Agostino et al., 2008).

Cependant, malgré cette simplicité opérationnelle, FRS-SP partage plusieurs limites avec FRS. Il tend à sous-estimer le risque chez les femmes (K. E. Park & Pepine, 2015), contribuant ainsi à des inégalités dans la prévention cardiovasculaire. Plus largement, les modèles de prédiction basés sur la régression de Cox présentent des performances variables selon les populations et peinent à capturer certaines interactions non linéaires ou des interactions entre facteurs de risque (Cho et al., 2021; Rajula et al., 2020). Ils présentent souvent des faiblesses en matière de discrimination (capacité à distinguer les cas à risque des non-cas) et d'étalonnage (précision des probabilités prédictives), ce qui réduit leur utilité clinique pour des prédictions (Hu et al., 2021).

Pour répondre à ces limites, des modèles reposant sur l'AA émergent comme un outil prometteur d'aide à la décision clinique (Arnould et al., 2023). Ils permettent d'analyser de larges volumes de données médicales et de mieux détecter les interactions non linéaires et offrent ainsi un potentiel d'amélioration de l'identification des personnes à haut risque (Gautam et al., 2023 ; Moncada-Torres et al., 2021 ; Premasagar et al., 2022). De plus, ils permettent d'identifier les facteurs de risque les plus influents dans la prédiction (Song et al., 2023). Ainsi, les algorithmes d'AA jouent un rôle important dans la prise de décision clinique, contribuant à une meilleure prédiction, stratification et planification des traitements (Darabi et al., 2024). Malgré ces avancées, plusieurs lacunes importantes persistent dans la littérature. Tout d'abord, de nombreuses approches d'AA utilisées pour la prédiction du risque cardiovasculaire ne tiennent pas pleinement compte de la nature temporelle des données. Ensuite, les analyses stratifiées selon le sexe restent encore limitées, malgré les différences bien documentées et le risque des MCV. Par ailleurs, peu de travaux proposent des comparaisons systématiques entre différents modèles AA de survie tels que le RSF, GBSA et DeepSurv.

1.2 OBJECTIFS

Pour bien gérer et apporter des réponses claires à notre problématique, nous avons fixé un objectif général et des objectifs spécifiques.

Objectif général :

L'objectif principal de cette étude est de développer et de comparer les performances prédictives de modèles d'AA de survie (RSF, GBSA et DeepSurv) pour la prédiction du risque cardiovasculaire à 10 ans au sein d'une cohorte de 6883 participants, en ayant recours aux prédicteurs d'un outil existant, l'outil de FRS-SP.

Une méthodologie rigoureuse visant à assurer la robustesse et la reproductibilité des résultats sera appliquée. Cette étude met en avant l'efficacité de l'utilisation des modèles AA de survie pour la prédiction du risque cardiovasculaire.

Objectifs spécifiques :

Afin de réussir notre objectif général, nous nous engageons à atteindre trois objectifs spécifiques :

- Il s'agira dans un premier temps de comparer trois modèles d'apprentissage automatique de survie : le RSF, le GBSA et le DeepSurv.
- Ensuite, l'étude vise à évaluer la performance des trois modèles (calibration et discrimination).
- Enfin, c'est d'analyser les différences de risques selon le sexe et les catégories de risque (faible : <10 %, modéré : 10-20 %, élevé : >20 %).

1.3 METHODOLOGIE

1.3.1 Devis

Le présent projet a utilisé des données de l'Étude prospective québécoise sur le travail et la santé, « The PROspective Quebec (PROQ) Study on Work and Health », ou Étude PROQ, cette cohorte a été décrite avec plus de détails par Trudel et al (Trudel et al., 2018). La cohorte initiale comprenait 9188 participants cols blancs, âgés de 18 à 65 ans employés dans 19 entreprises publiques de la région de Québec ont été recrutés entre 1991 et 1993 (T1), avec un taux de participation de 75 %. L'étude PROQ comporte deux suivis : T2 (1999-2001) et T3 (2015-2018) avec des taux de participation respectifs de 89 % et 81 % les personnes éligibles. De T1 à T3, la durée moyenne de suivi était de 24 ans, tandis qu'elle était de 18 ans entre T2 et T3. L'accès aux bases de données médico-administratives était disponible pour 96 % de la cohorte initiale.

1.3.2 Population de l'étude

Cette étude inclut 6883 travailleurs cols blancs, sans maladie cardiovasculaire, provenant de la cohorte PROQ composante cardiovasculaire, présents lors de la collecte de données de 1999 à 2001. Les participants ayant des antécédents de MCV à ce moment, ceux dont l'âge n'était pas compris entre 30 à 74 ans et ceux présentant des données manquantes sur les prédicteurs ont été exclus. Après ces exclusions, un échantillon analytique final de 6 883 participants (âge moyen : 47 ans ; 50.5% de femmes) restait éligible sur une période de suivi de 10 ans. Comme dans la plupart des cohortes populationnelles, la prévalence des MCV est faible, ce qui entraîne un déséquilibre marqué entre les classes. Au cours du suivi de 10 ans, **9.1 %** des participants ont présenté un événement cardiovasculaire et **90.9 %** n'ont pas présenté d'événement.

1.3.3 Variables à l'étude

1.3.3.1 Facteurs de risque de Framingham

Les variables considérées dans la présente étude, sont les facteurs de risques cardiovasculaires de l'outil simplifié pour soins primaires (Framingham-SP). Elles provenaient du questionnaire auto-administré de l'étude PROQ. Les facteurs de risque sont :

- **Variables sociodémographiques** : L'âge est traité comme une variable continue, et le sexe comme une variable catégorielle (1: homme, 2: femme).
- **Comorbidités** : La médication pour hypertension a été considérée comme une variable binaire (oui/non) « Est-ce qu'un médecin, une infirmière ou un autre professionnel de la santé vous a déjà dit que votre tension artérielle était trop élevée (ou que vous faisiez de la haute pression)? Le diabète (oui/non). « Est-ce qu'un médecin, une infirmière ou un autre professionnel de la santé vous a déjà dit que vous faisiez du diabète? »
- **Habitudes de vie** : Le tabagisme a été traité comme une variable binaire (fumeur actuel: oui/non) et évalué sur sept éléments: « Fumez-vous de la cigarette? » « À quel âge avez-vous commencé à fumer tous les jours? » ; « Combien de cigarettes fumez-vous habituellement par jour? » ; « Avez-vous déjà fumé la cigarette tous les jours? » ; « À quel âge avez-vous commencé à fumer la cigarette tous les jours? » ; « À quel âge avez-vous cessé de fumer la cigarette tous les jours? » ; « Environ combien de cigarettes fumiez-vous habituellement par jour? ».
- **Mesures anthropométriques et cliniques** : L'indice de masse corporelle (IMC), substituant le cholestérol dans ce modèle, a été calculée en divisant le poids par la taille au carré (kg/m^2) et catégorisée en poids normal ($<25 \text{ kg/m}^2$), excès de poids ($25\text{-}29.9 \text{ kg/m}^2$) et obésité ($\geq 30 \text{ kg/m}^2$), selon des protocoles validés. La pression artérielle systolique est mesurée au repos avec un sphygmomanomètre, conformément aux recommandations de l'American Heart Association (Whelton et al., 2018), sur les deux bras, après une période de repos de cinq minutes en position assise.

Les moyennes de deux mesures de la pression artérielle prises au départ, espacées de 1 à 2 minutes, ont été utilisées comme niveaux de pression artérielle.

- Enfin, le risque de MCV a été catégorisé en trois niveaux : faible ($<10\%$), modéré ($10-19\%$) et élevé ($\geq 20\%$), conformément aux seuils recommandés par la Société canadienne de cardiologie (Pearson et al., 2021).

1.3.3.2 Incidence des maladies cardiovasculaires

L'incidence des maladies cardiovasculaires (MCV) a été définie comme la survenue d'un premier événement cardiovasculaire, incluant les maladies coronariennes, les accidents vasculaires cérébraux (AVC), la maladie artérielle périphérique athérosclérotique des membres inférieurs (MAP) ou de l'insuffisance cardiaque congestive (ICC). L'ensemble des cas survenus après le début de l'étude (T2) ont été inclus pour toute la durée du suivi, l'événement MCV est défini de manière binaire : présence (1) ou absence (0) d'un événement survenu après T2.

Les codes diagnostiques ont été tirés de la Classification internationale des maladies (CIM), tandis que les codes de procédure provenaient de la Classification canadienne des interventions en santé (CCI) et de la Classification canadienne des actes diagnostiques, thérapeutiques et chirurgicaux (CCP). Des informations sur les causes de décès ont également été utilisées. Plus précisément, les codes diagnostiques issus de la neuvième révision (CIM-9) et de la dixième révision (CIM-10), ainsi que les codes d'intervention (CCP et CCI), ont été extraits des fichiers des services médicaux rémunérés à l'acte de la Régie de l'assurance maladie du Québec (RAMQ) et du fichier des hospitalisations du Québec.

Les cas de maladies coronariennes ont été identifiés avec une sensibilité de 77,0 %, une spécificité de 97,5 % et une valeur prédictive positive égale à 75,3 %. Cela inclut l'infarctus du myocarde, l'angine, ainsi que le syndrome coronarien aigu et chronique (codes CIM-9 : 410-414 ; CIM-10 : I20-I25), de même que les cas de revascularisation coronarienne, qu'il s'agisse d'angioplastie ou de chirurgie de pontage (codes CCP : 48.02, 48.03, 48.09, 481 et codes CCI : 1.IJ.50, 1. IJ.57.GQ, 1.IJ.76) (Tu et al., 2010).

L'algorithme, affichant une sensibilité de 84,8 %, une spécificité de 97,0 % et une valeur prédictive positive de 55,6 % a permis d'identifier les individus atteints d'insuffisance cardiaque congestive (ICC) à partir des codes CIM-9 428 ou CIM-10 I500, I501 et I509 lors d'une hospitalisation ou de soins ambulatoires, suivi d'un autre épisode de soins dans l'année. Deux définitions sont envisagées : une définition stricte, basée uniquement sur les codes CIM-9 428 et CIM-10 I500, I501 et I509 et une définition élargie, intégrant également les codes associés à la myocardiopathie (CIM-9 : 425, CIM-10 : I42) et à l'œdème pulmonaire (CIM-9 : 514, 518.4, CIM-10 : J81) (Schultz et al., 2013).

L'algorithme de diagnostic de la maladie artérielle périphérique athérosclérotique des membres inférieurs, qui utilise les codes suivants : ICD-9 (440.20-440.24, 440.9, 443.9), CIM-10 (I70.2, I70.9, I73.9) présente une sensibilité de 86 %, une spécificité de 83 % et une valeur prédictive positive de 91 %. Ces résultats démontrent une performance robuste dans l'identification artérielle périphérique athérosclérotique des membres inférieurs à partir des données médico-administratives (Fan et al., 2013).

1.3.4 Analyses

L'approche que nous avons proposée pour la prédiction du risque cardiovasculaire est structurée en six (6) étapes clés, chacun jouant un rôle essentiel dans le processus de la modélisation et l'évaluation des performances des modèles.

Les principales étapes méthodologiques sont les suivantes : 1) Prétraitement et division des données, 2) Optimisation des hyperparamètres, 3) Gestion du déséquilibre des données, 4) Développement des modèles, 5) Mesures de performances des modèles et enfin 6) Importances des variables. La figure méthodologique proposée est illustrée à la Figure 2.

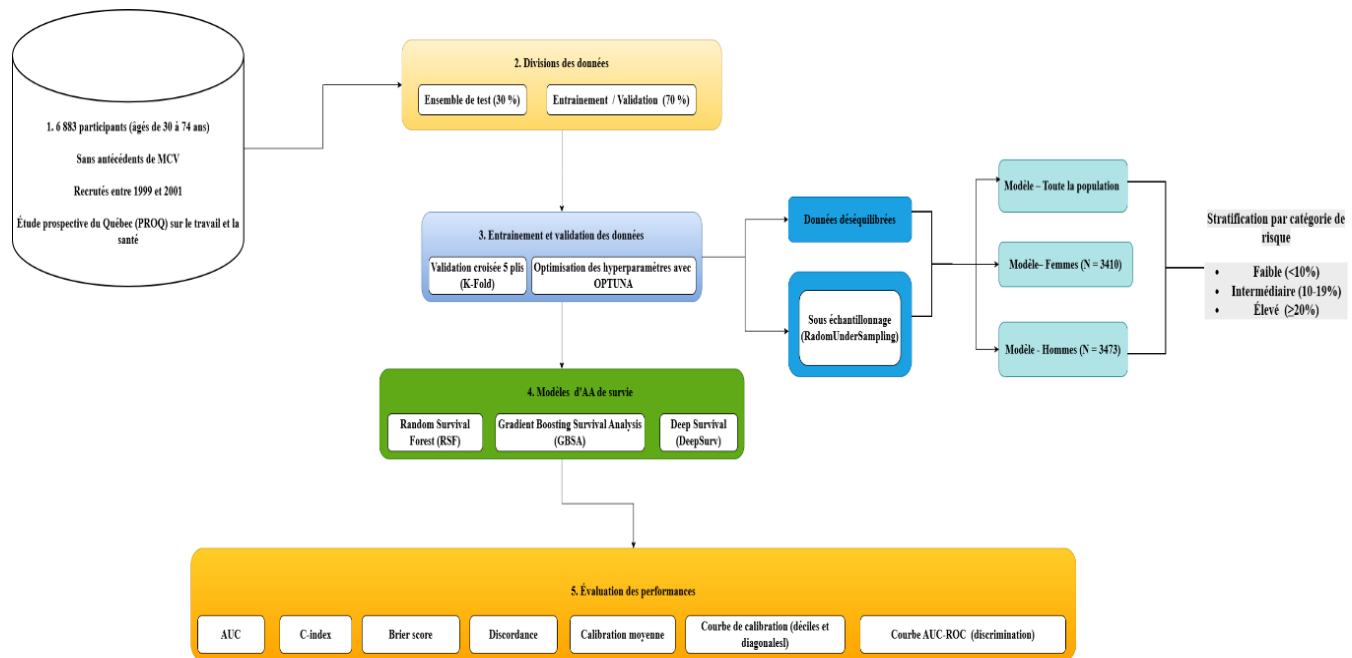


Figure 1. *Méthodologie proposée pour une prédiction du risque MCV*

1.3.4.1 Prétraitement et division des données

D'abord, les données initiales ont été reçues dans un format déjà largement nettoyé et prétraité. Celui-ci inclut le traitement de valeurs manquantes et une sélection des variables. Ensuite, l'ensemble des données a été divisé en trois parties : entraînement, validation et test, 70% pour l'entraînement et la validation et 30% pour le test afin d'évaluer la performance sur de données non encore vues par les modèles (Singh et al., 2021). L'entraînement a été effectué avec une validation croisée K-Fold de 5 plis pour permettre d'éviter le sur apprentissage des modèles (Bradshaw et al., 2023) et d'estimer l'erreur de prédiction des modèles dans l'entraînement et dans le test (Rodríguez et al., 2010). Après prétraitement, les modèles ont été ajustés à partir des données d'entraînement, puis leur validité a été vérifiée à l'aide des données de test.

1.3.4.2 Optimisation des hyperparamètres

Le choix des valeurs des hyper paramètres dépend de la complexité des données, de la nécessité d'éviter le sur apprentissage (Shekhar et al., 2022). Dans cette étude, nous avons utilisé Optuna, une méthode de nouvelle génération basée sur l'optimisation bayésienne pour trouver les meilleurs hyper paramètres (Akiba et al., 2019). Elle a été combinée avec la validation croisée à 5 plis, afin de sélectionner les combinaisons d'hyper paramètres les plus performantes pour les modèles.

1.3.4.3 Déséquilibre des classes

Dans cette étude, le jeu de données présente un déséquilibre important entre les classes, la classe majoritaire (individus sans MCV : 90.9%) étant numériquement plus représentée que la classe minoritaire (individus avec MCV : 9.1%). Ceci est illustré dans Figure 2. Ce déséquilibre est susceptible d'induire un biais d'apprentissage dans les modèles, les amenant à privilégier la classe majoritaire et à réduire leur performance prédictive sur la classe minoritaire (F. Zhang et al., 2022).

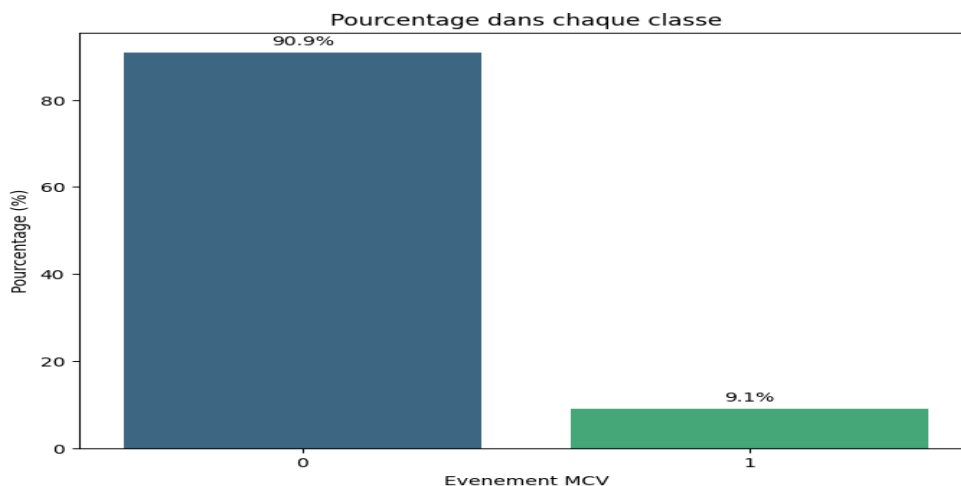


Figure 2. *Distribution des classes*

Afin de mitiger l'impact de ce déséquilibre, deux stratégies distinctes ont été explorées : l'optimisation des hyperparamètres et le sous-échantillonnage de la classe majoritaire.

a. Optimisation des hyperparamètres

Une première approche a consisté à intégrer l'ajustement des hyperparamètres lors de l'entraînement. Des travaux ont montré que certains hyperparamètres jouent un rôle clé dans la prise en compte des événements rares. L'augmentation du nombre d'arbres (*n_estimators*) améliore la stabilité des prédictions et augmente la probabilité de capturer correctement des événements rares (Ishwaran et al., 2008). De même, la limitation du nombre de variables testées à chaque division (*max_features*) contribue à éviter que les données majoritaires dominent systématiquement les splits (Wright & Ziegler, 2017). Enfin, des paramètres tels que la taille minimale des feuilles (*min_samples_leaf*) ou le nombre minimal d'observations pour diviser un nœud (*min_samples_split*) permettent de préserver la présence de la classe minoritaire dans les partitions (Ishwaran & Lu, 2019). Cette stratégie a l'avantage de conserver l'intégralité du volume de données initial sans nécessiter le retrait d'instances de la classe majoritaire.

b. Sous-échantillonnage de la classe majoritaire

La deuxième stratégie a impliqué l'application d'un sous-échantillonnage (*undersampling*) sur la classe majoritaire au sein de l'ensemble d'entraînement. L'objectif était d'obtenir une distribution plus équilibrée entre les classes pour l'apprentissage, limitant ainsi l'influence prépondérante de la classe majoritaire sur l'entraînement du modèle (Yousefimehr et al., 2025). Ceci est illustrée dans la Figure 3.

Dans les deux cas, les performances des modèles ont été évaluées et comparées sur un même ensemble de test indépendant, s'assurant qu'il n'avait jamais été exposé aux modèles durant la phase d'entraînement, afin de garantir une comparaison objective et équitable des deux méthodes.

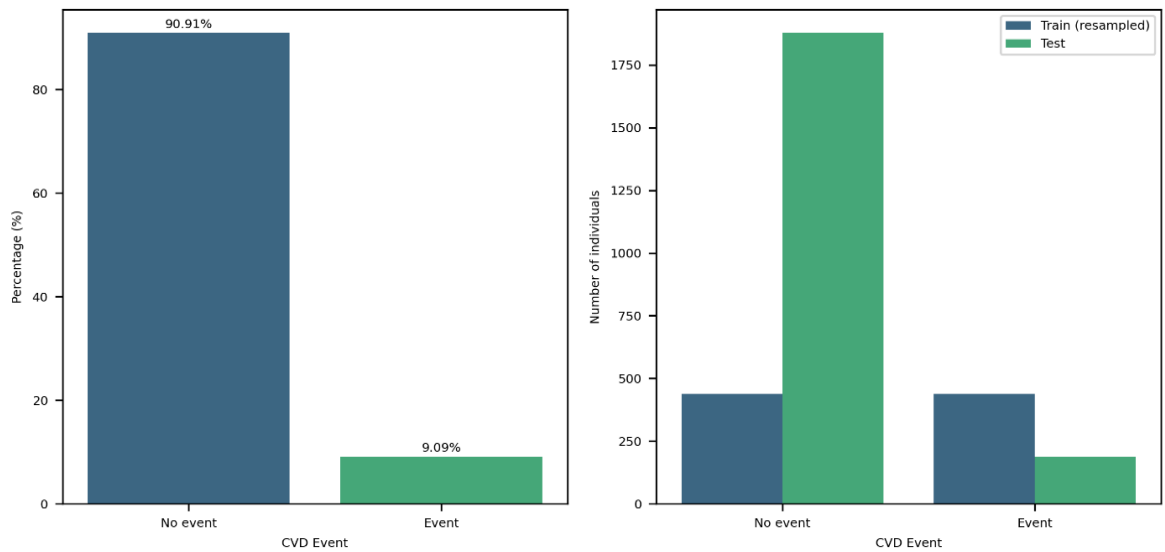


Figure 3. *Distribution des classes avant et après le sous-échantillonnage*

1.3.4.4 Développement des modèles

Le risque moyen prédit de MCV à 10 ans a été estimé sur la base des variables du FRS-SP à l'aide de trois modèles d'apprentissage automatique pour l'analyse de survie : RSF, GBSA et DeepSurv. Ces modèles ont été sélectionnés en raison de leurs performances rapportées dans des études antérieures (Bao et al., 2022; Blanchard et al., 2022; Chun et al., 2021; Sabovčik et al., 2022) et de leur utilisation fréquente pour la prédiction du risque cardiovasculaire (Teshale et al., 2024).

Pour chaque modèle, le risque prédit a été calculé à partir de sa fonction de risque cumulatif estimée. Le risque observé à 10 ans a été obtenu en utilisant la fonction d'incidence cumulative. Le développement des modèles a été réalisé à l'aide du langage Python (version 3.11.7, 64 bits), dans l'environnement Jupyter Notebook (version 7.0.8), en s'appuyant sur des bibliothèques open source dédiées à l'AA et à l'analyse de survie, telles que *scikit-survival*, *PySurvival*, *lifelines* et *PyTorch* (notamment pour le modèle DeepSurv).

1.3.4.5 Mesures de performances des modèles

La calibration et la discrimination ont été utilisés, tel que recommandé par Van Calster et al (Van Calster et al., 2019) et Steyerberg et al (Steyerberg, 2008) pour évaluer la performance des modèles de prédiction. La calibration permet de mesurer la concordance entre les risques prédits et les risques observés (Austin & Steyerberg, 2019). La calibration moyenne est estimée comme le rapport entre le risque moyen prédit et le risque observé sur la période de suivi. Elle est acceptable si la valeur est comprise entre 0.9 et 1.1, les valeurs inférieures à cet intervalle indiquent une sous-estimation du risque, tandis que des valeurs supérieures montrent une surestimation (Riley et al., 2021).

La discordance est une mesure de calibration, calculée à l'aide de la formule suivante :

$$\text{Discordance} = \frac{\text{Risque prédit} - \text{Risque observé}}{\text{Risque observé}} \times 100$$

Une discordance de 0 indique une calibration parfaite ; une valeur négative indique une sous-estimation, tandis qu'une valeur positive indique une surestimation du risque (Austin & Steyerberg, 2019).

La discrimination évalue la capacité du modèle à distinguer les individus à haut risque de ceux à faible risque. Le risque prédit sur 10 ans a été évalué à l'aide de deux mesures de performance : l'indice C (également appelé C-index ou C-statistics) qui quantifie la capacité du modèle à différencier les individus ayant présenté un événement cardiovasculaire de ceux n'en ayant pas présenté (S. Y. Park et al., 2021; Van Geloven et al., 2022). L'aire sous la courbe ROC (Receiver Operating Characteristic), largement utilisée pour évaluer la capacité discriminante d'un modèle. La ligne pointillée représente une AUC de 0,5 (soit une prédiction aléatoire). Plus la courbe ROC s'éloigne de cette ligne et se rapproche du coin supérieur gauche, plus la capacité discriminante du modèle est jugée satisfaisante (D'Agostino & Nam, 2003).

Le score de Brier a été utilisé comme mesure globale de performance. Il reflète à la fois la calibration et la discrimination. Ce score correspond à la moyenne des carrés des écarts entre les probabilités prédites et les événements observés. Le score de Brier varie de 0 à 1.

Des valeurs supérieures à 0,25 indiquent une faible performance du modèle, tandis que des valeurs inférieures à 0,25 indiquent une meilleure performance de prédiction (Steyerberg et al., 2010). Des analyses ont été réalisées à la fois dans la population totale et de manière stratifiée selon le sexe (hommes et femmes). Les risques prédits et observés ont été comparés dans les différentes catégories de risque (faible : <10 %, modéré : 10-19 %, élevé : ≥ 20 %).

1.3.4.6 Importance des variables

Pour évaluer la capacité prédictive de chaque facteur de risque sur la prédiction du risque cardiovasculaire, nous avons procédé à une analyse de l'importance des variables pour chacun des trois modèles. Pour les modèles RSF et GBSA, l'importance des variables a été calculée à partir de la diminution de l'impureté (mean decrease in impurity) au niveau des arbres. Cette mesure reflète la contribution relative de chaque variable à la réduction de l'erreur de prédiction dans l'ensemble du modèle. Pour le modèle RSF, nous avons utilisé l'importance par permutation pour calculer l'importance des variables (Zhao et al., 2025).

Considérations éthiques

Ce projet s'intégrait dans un programme plus large de recherche déjà approuvé par le comité d'éthique du Centre de recherche du CHU de Québec-Université Laval. Les participants qui ont pris part à l'étude de cohorte PROQ sur le travail et la santé ont fourni un consentement écrit, libre et éclairé. Des assistants de recherche les ont rencontrés individuellement pour leur expliquer le contexte, les objectifs ainsi que le déroulement de l'étude. Un participant pouvait abandonner le projet à tout moment sans fournir d'explications. La confidentialité et l'anonymat des données ont été respectés grâce à l'attribution d'un numéro d'identification à chaque participant ; ce qui les rendait non identifiables sur le fichier principal. Les informations nominales et les numéros d'identification ont été gardés dans un autre fichier confidentiel auquel seuls les chercheurs principaux et le personnel assigné à la gestion de ces informations ont eu accès. Enfin, on doit souligner que ce projet ne représentait aucun risque pour la santé des participants.

1.4 CONTRIBUTIONS

Face aux limites identifiées dans la littérature, notre étude propose une méthodologie structurée visant à améliorer la robustesse et la pertinence clinique des modèles de prédiction du risque cardiovasculaire. Nos contributions s'articulent autour des éléments suivants :

- Nous procédons à une comparaison simultanée des approches RSF, DeepSurv et GBSA sur un même jeu de données, combinées avec les prédicteurs de FRS-SP, contrairement aux études antérieures qui évaluent les modèles de manière isolée. Nous proposons également une méthodologie complète incluant la préparation des données, la validation croisée et l'optimisation des hyper paramètres. Face à l'absence d'analyse approfondie du déséquilibre des données dans la littérature existante, nous évaluons aussi deux approches distinctes : l'ajustement des hyperparamètres pour atténuer le déséquilibre et l'application d'un sous-échantillonnage (Under-sampling) de la classe majoritaire.
- Alors que la majorité des études se limitent aux mesures de discrimination (C-index/AUC). Nous intégrons également une analyse détaillée de la calibration via le score de Brier et de courbes de calibration basées sur des déciles et une approche diagonale. Cette analyse vise à apporter un éclairage supplémentaire sur la fiabilité clinique des prédictions. Contrairement à la recherche existante, souvent concentrée quasi exclusivement sur les métriques de performance globales, nous évaluons les modèles selon les catégories de risque distinctes (faible, modéré, élevé), mais également selon le sexe. Cette double analyse est cruciale pour évaluer la pertinence réelle des modèles à travers des sous-populations clés. Enfin, les résultats obtenus ont conduit à la rédaction d'un article scientifique actuellement finalisé et destiné à une soumission à la revue *JMIR (Journal of Medical Internet Research)* plus particulièrement à *JMIR Cardio*.

1.5 STRUCTURE DU MÉMOIRE

Ce mémoire structuré en deux chapitres adopte le format d'un mémoire par articles, encadrés par une introduction générale et une conclusion générale.

L'introduction générale présente une vue d'ensemble du domaine de la prédiction du risque des maladies cardiovasculaires et des modèles d'AA appliqués aux données de survie. Elle met en évidence les limites des approches traditionnelles et justifie l'intérêt de recourir à des modèles d'AA pour prédire le risque cardiovasculaire en soulignant le contexte et la problématique. L'établissement des objectifs de l'étude, la méthodologie de travail et la contribution terminent ce premier segment.

Le premier chapitre présente les concepts fondamentaux et une revue de la littérature sur les études antérieures. Il introduit d'abord les notions clés liées aux MCV, notamment leur définition, leurs principaux facteurs de risque ainsi que leur épidémiologie. Il aborde également le concept de risque cardiovasculaire et son importance dans la prévention et la prise en charge clinique. Il décrit ensuite les principaux modèles utilisés, allant des approches statistiques traditionnelles jusqu'aux modèles d'AA de survie appliquées à la prédiction du risque cardiovasculaire. Ensuite, les principes de l'analyse de survie sont présentés, en mettant l'accent sur la gestion de la censure et la dimension temporelle des données. Enfin, une synthèse critique des travaux existants est proposée afin d'identifier les limites actuelles et de motiver la problématique de recherche.

Le deuxième chapitre présente l'article scientifique, qui a été finalisé et est prêt pour soumission au Journal JMIR Cardio (Journal of Medical Internet Research), une revue scientifique en libre accès (Open Access). Cet article porte sur la prédiction des MCV sur 10 ans à l'aide de modèles d'AA adaptés à l'analyse de survie, appliqués à l'outil de FRS-SP. Il repose sur l'utilisation de trois modèles AA de survie afin d'évaluer et de comparer leurs performances dans l'estimation du risque cardiovasculaire.

Le mémoire se termine par une conclusion générale de l'étude en présentant un résumé des principaux résultats, en discutant des implications et des limites et des perspectives pour les travaux futurs.

CHAPITRE 1 : REVUE DE LA LITTERATURE

Ce chapitre présente deux sections : les principaux concepts nécessaires à la compréhension de la problématique de recherche ainsi qu'une revue de la littérature sur la prédiction du risque cardiovasculaire portant sur les modèles d'AA de survie.

La première section est divisée en deux parties : une première partie consacrée aux fondements théoriques relatifs aux MCV, incluant leur définition, leur épidémiologie, ainsi que leurs facteurs de risque. Une deuxième partie présentant les approches de prédiction du risque cardiovasculaire, en distinguant les modèles traditionnels fondés sur la régression statistique et les modèles d'AA de survie utilisés dans le cadre de ce mémoire (RSF, GBSA et DeepSurv).

Dans la deuxième section, on a aussi deux parties : une première consacrée à la stratégie de recherche documentaire et une deuxième, présentant une synthèse critique de la littérature scientifique, en mettant en lumière des études ayant évalué ou comparé les performances des modèles d'AA de survie pour la prédiction des événements cardiovasculaires.

Cette démarche permettra de situer la présente étude dans le contexte scientifique actuel et de mettre en évidence les contributions attendues du travail de recherche.

1.6 CONCEPTS FONDAMENTAUX

1.6.1 Définition des maladies cardiovasculaires

Les MCV regroupent un ensemble de pathologies affectant le cœur et les vaisseaux sanguins. Elles incluent des affections telles que la cardiopathie ischémique, l'accident vasculaire cérébral (AVC), l'accident ischémique transitoire (AIT), l'insuffisance cardiaque congestive (ICC) ainsi que la maladie artérielle périphérique (MAP) (OMS, 2022).

1.6.2 Épidémiologie des maladies cardiovasculaires

En raison de leur forte incidence, de la mortalité associée et des conséquences fonctionnelles à long terme, les MCV génèrent un fardeau épidémiologique majeur, tant pour les individus que pour les systèmes de santé (Darabi et al., 2024). Le nombre de décès dus aux MCV a augmenté, passant de 12,3 millions en 1990 à plus de 17,6 millions en 2016, faisant de ces pathologies la principale cause de mortalité dans le monde (Meier et al., 2019). Les MCV génèrent entre 7 et 21 % des coûts de santé mondiaux, englobant à la fois les dépenses directes et indirectes (Santos et al., 2020).

Au Canada, les maladies cardiaques, notamment les MCV représentent la deuxième cause de décès et l'une des causes prédominantes d'hospitalisation au cours des 20 dernières années représentant 24 % de tous les décès en 2020 avec un fardeau économique estimé à 13 milliards CAD en 2010 (Gilbert-Ouimet et al., 2024). Une étude basée sur une cohorte observationnelle du registre REACH au Canada rapporte aussi que le coût moyen annuel d'hospitalisation est d'environ 1 743 \$ CAN pour la maladie coronarienne, 1 823 \$ CAN pour une maladie cérébrovasculaire, et 4 677 \$ CAN pour une maladie artérielle périphérique (Smolderen et al., 2010).

Ces chiffres illustrent le fardeau considérable des MCV, tant sur le plan sanitaire que socio-économique, et justifient l'importance des stratégies de prévention et des outils de prédiction du risque cardiovasculaire

1.6.3 Risque cardiovasculaire

Un risque cardiovasculaire correspond à la probabilité qu'un individu développe une maladie cardiovasculaire ou un événement cardiovasculaire afin de soutenir la prédiction et la prise en charge clinique des individus, à partir d'un ensemble de facteurs de risque identifiés (Thiriet, 2019). L'estimation du risque cardiovasculaire permet d'identifier les individus les plus susceptibles de développer une MCV dans une période de temps donnée (Grover & Lowensteyn, 2011).

1.6.4 Facteurs de risque cardiovasculaires

Les MCV augmentent avec l'âge et que la plupart des décès surviennent chez les personnes âgées (Baudin et al., 2009; Jain et al., 2023; Y. Zhang et al., 2020; Zores & Maeker, 2022). Cependant d'une manière générale, le risque de développement d'une MCV donnée peut varier en fonction de la présence ou de l'absence de caractéristiques individuelles. L'hypertension artérielle est considérée comme étant le principal facteur de risque MCV, pratiquer un dépistage selon l'âge et le sexe, plus précocement pour les familles et sujets à haut risque peut être une bonne solution. Néanmoins pour (Jegan & Nimi, 2023), l'inactivité physique est considérée comme le quatrième principal facteur de risque de mortalité mondiale associée à 6% des décès dans le monde, et que la promotion de l'activité physique peut réduire le risque de la survenue des MCV. En effet, pour (Artigao-Rodenas et al., 2013), il est essentiel d'utiliser les quatre principaux facteurs de risque ayant le plus de poids, comme reconnus en épidémiologie cardiovasculaire : le cholestérol, la pression artérielle, le diabète sucré et le tabagisme. Ces dernières années, la découverte de nouveaux facteurs de risque cardiovasculaire a donné l'espoir de mieux évaluer le risque global (Blacher & Agnoletti, 2011). La prévention des MCV doit se baser sur l'évaluation du risque individuel des personnes et la connaissance du niveau du risque doit considérer la totalité des facteurs de risque et les caractéristiques personnelles de chaque personne.

1.6.5 Prédiction du risque cardiovasculaire

Cette partie a pour but de rapporter les données issues de la littérature en ce qui concerne la prédiction du risque des MCV. Dans ce contexte, il est essentiel d'avoir des modèles pour prédire le risque cardiovasculaire en fonction des différents facteurs identifiés et du type de sortie prédite attendue.

Trois types de sorties prédictives peuvent être distinguées :

- **La classification binaire** : prédire la présence ou l'absence d'un événement cardiovasculaire.
- **La classification multiple** : prédire le niveau de risque cardiovasculaire selon plusieurs catégories (faible, modéré, élevé).
- **La régression pour la prédiction du risque absolu et temporel** : estimer un score de risque ou prédire la survenue d'un événement dans un intervalle de temps donné (modèles de survie).

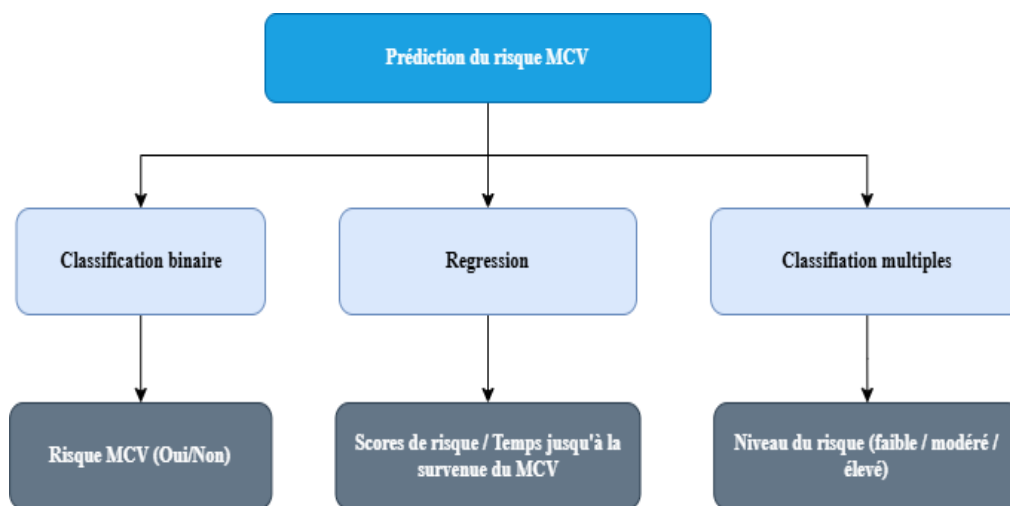


Figure 4. Prédiction du risque

1.6.5.1 Modèles de risque traditionnels

Scores de prédiction du risque cardiovasculaire

Il est important de disposer des modèles de prédiction de risque permettant de prédire la survenue d'un nouvel événement cardiovasculaire (Lloyd-Jones, 2010). Le dépistage à long terme des MCV, l'évaluation du risque et l'initiation de plans de prévention reposent généralement sur des modèles de prédictions de risques conventionnels.

- **Le SCORE** (Systematic Coronary Risk Evaluation)

Il est développé à partir de 12 cohortes européennes et prédit le risque de mortalité cardiovasculaire sur 10 ans (Conroy et al., 2003).

- **Le calculateur d'équations de cohorte regroupées (PCE),**

Il prédit le risque sur 10 ans d'événements coronariens sévères (infarctus du myocarde), de décès par coronaropathie et d'accidents vasculaires cérébraux (AVC). Il a été développé dans des cohortes de 5 populations blanches et afro-américaines aux États-Unis (Preiss & Kristensen, 2015).

- **Le score de Suita**

Il est un modèle japonais de prédiction des maladies coronariennes avec une cohorte prospective évaluant les cardiopathies congénitales d'apparition récente dans la ville de Suita, à Osaka, au Japon (Hirai et al., 2019).

- **Le QRISK**

C'est un outil spécifiquement conçu pour la population du Royaume-Uni, permettant de prédire le risque à 10 ans d'événements cardiovasculaires (Hippisley-Cox et al., 2007).

- **Le score Assign**

Dérivé du score de Framingham, en plus des facteurs de risques de Framingham, considère la déprivation sociale et les antécédents familiaux de maladies cardiovasculaires comme des facteurs de risque significatifs (Woodward et al., 2007).

- **Le score de Framingham (FRS)**

Il est le fruit d'une étude longitudinale menée dans la ville du même nom au Massachusetts (États-Unis) (Dawber et al., 1963). Développé à partir de la Framingham Heart Study, une

étude de cohorte prospective débutée en 1948 aux États-Unis et financée par le *National Institutes of Health* puis gérée par l'Université de Boston (Andersson et al., 2021). Un pourcentage inférieur à 10% est associé à un risque faible, un pourcentage entre 10 et 19% est associé à un risque modéré et un pourcentage supérieur ou égal à 20% est considéré comme étant un risque élevé. Ce sont les seuils recommandés par la Société Canadienne de Cardiologie (Pearson et al., 2021).

1.6.5.2 Modèles d'apprentissage automatique de survie

Définition de l'apprentissage automatique

L'AA ou aussi appelé Machine Learning (ML), est un domaine scientifique et plus particulièrement un sous-domaine de l'Intelligence Artificielle (IA) qui permet aux ordinateurs d'acquérir des capacités d'apprentissage (Sah, 2020). L'AA vise principalement à donner aux machines la capacité d'apprendre à partir des données, en repérant des modèles et en les généralisant afin de prendre des décisions ou de prédire de nouvelles données (Sarker, 2021). Elle englobe diverses méthodes permettant la création automatique de modèles à partir de données afin de prendre des décisions ou de prédire de nouvelles données. Ces méthodes se manifestent sous la forme d'algorithmes conçus pour analyser et interpréter les données, facilitant ainsi l'apprentissage autonome des systèmes informatiques (Mahesh, 2019).

Types d'apprentissage automatique

Les algorithmes d'AA peuvent être classés en quatre types : apprentissage supervisé, non supervisé, semi-supervisé et par renforcement (Figure 5).

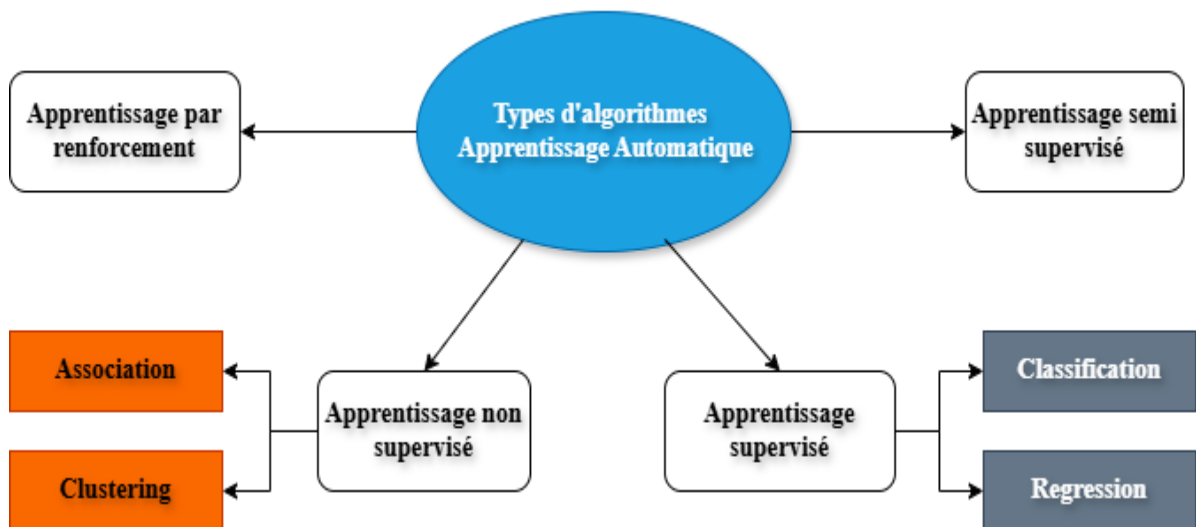


Figure 5. *Types d'Apprentissage Automatique*

- **Apprentissage supervisé**

L'apprentissage automatique supervisé est la construction des algorithmes qui consiste à créer des modèles et des hypothèses générales en utilisant des instances externes afin de prédire les valeurs de sorties (Singh et al., 2023).

- **Apprentissage non supervisé**

L'apprentissage non supervisé est une technique d'apprentissage automatique où un modèle est élaboré à partir d'un ensemble de données d'entraînement restreint aux entrées. L'apprentissage du modèle consiste à repérer les structures et les modèles présents dans les données sans nécessiter de sorties étiquetées (Pietrobon et al., 2025) .

- **Apprentissage semi-supervisé**

L'apprentissage semi supervisé consiste à utiliser des données non-annotées afin de compléter l'apprentissage supervisé. Il est spécialement conçu pour les secteurs d'application où les données non étiquetées sont fréquentes, comme le traitement des images, la collecte d'informations et la bio-informatique (van Engelen & Hoos, 2020; Y C a et al., 2018).

- **Apprentissage par renforcement**

L'apprentissage par renforcement est appliqué dans des domaines tels que les jeux, la navigation et la robotique. Son mécanisme repose sur la détermination de l'action optimale. L'agent agit dans son environnement, reçoit des retours sous forme de récompenses ou de pénalités, et adapte sa stratégie pour maximiser la récompense cumulée au fil du temps (Pietrobon et al., 2025).

Apprentissage automatique appliqué à l'analyse de survie

- **Analyse de survie**

L'analyse de survie est un domaine visant à analyser et à modéliser des données pour lesquelles la variable d'intérêt correspond au temps écoulé jusqu'à la survenue d'un événement (P. Wang et al., 2017). Cette approche vise à estimer le temps jusqu'à la survenue d'un événement. Le modèle de Cox est le plus utilisé pour l'analyse multivariée en survie. Dans ce cadre, chaque individu est décrit par un indicateur d'événement et une durée de suivi. Lorsque l'événement n'est pas observé au cours du suivi, l'observation est considérée comme censurée, indiquant que l'événement pourrait survenir ultérieurement. (Blanchard et al., 2022). L'AA de survie regroupe des méthodes conçues pour modéliser le temps jusqu'à la survenue d'un événement tout en tenant explicitement compte de la censure (lorsque l'événement ne survient pas pendant la période de suivi ou que l'individu est perdu de vue), une limitation majeure des algorithmes d'AA classiques (Ishwaran et al., 2008). Ces approches permettent l'estimation de quantités telles que la fonction de survie, le risque instantané ou le risque cumulé, tout en bénéficiant de la flexibilité des modèles d'AA

(Katzman et al., 2018). Dans ce contexte, plusieurs modèles d'AA de survie ont été proposés afin d'estimer le risque temporel tout en tenant compte de la censure. Parmi ces approches d'apprentissage supervisé adaptées à la survie, ce mémoire se concentre sur trois modèles : Random Survival Forest (RSF), Gradient Boosting Survival Analysis (GBSA) et DeepSurv. Ces différents modèles partagent un objectif commun : modéliser le risque en tenant compte explicitement du temps jusqu'à l'événement et de la censure, ce qui constitue un élément central de l'analyse de survie.

- **Random Survival Forest (RSF)**

RSF est un modèle d'ensemble non paramétrique et une extension du modèle des forêts aléatoires (Ishwaran & Lu, 2019) adapté aux données de survie.

Les modèles RSF constituent une approche d'apprentissage ensembliste largement utilisée en analyse de survie. Elles ont été appliquées avec succès dans de nombreux contextes et montrent, dans plusieurs études, des performances supérieures à celles des modèles de survie traditionnels (Miao et al., 2015). Il repose sur l'agrégation de plusieurs arbres de décision pour estimer les risques de survie et peut améliorer considérablement la performance des prédictions (Romano et al., 2024; Ruysinck et al., 2016).

Les différentes étapes de l'estimation du risque avec RSF :

- 1) **Échantillonnage bootstrap** : tirer $n_{\text{estimators}}$ (le nombre d'arbres dans la forêt) échantillons bootstrap à partir des données originales. Chaque échantillon bootstrap exclut en moyenne 37% des données (provient d'une propriété statistique du bootstrap sampling) appelées données out-of-bag (OOB).
- 2) **Sélection des variables** : à chaque nœud de l'arbre, sélectionner aléatoirement des variables candidates. Le nœud est divisé en utilisant la variable candidate qui maximise la différence de survie entre les nœuds fils.
- 3) **Prédiction par l'ensemble des arbres** : une fois tous les arbres construits, chaque arbre produit une fonction de survie individuelle pour un nouvel individu. La prédiction finale est obtenue en moyennant les fonctions de survie de tous les arbres de la forêt, ce qui permet d'obtenir une estimation plus robuste.

- 4) **Calcul de la fonction de risque cumulée (CHF)** : calculer une fonction de risque cumulée (*cumulative hazard function* (CHF)) pour chaque arbre, puis faire la moyenne pour obtenir la CHF d'ensemble.
- 5) **Estimation de la CHF avec les données OOB** : utiliser les données out-of-bag pour estimer une CHF d'ensemble, en moyennant les informations provenant des arbres éligibles. Ainsi, une estimation est obtenue pour chaque individu dans les données.
- 6) **Calcul du taux d'erreur de prédiction** : utilisation des données OOB pour calculer le taux d'erreur **de prédiction avec l'indice de concordance** (*C-index*).
- 7) **Deux différentes sorties** : fonction de survie estimée (probabilité que l'événement ne se produise pas avant un certain temps) et le risque cumulé (représente la probabilité que l'événement se produise à n'importe quel moment au fil du temps).

- **Gradient Boosting Survival Analysis (GBSA)**

GBSA est une méthode d'apprentissage ensembliste non paramétrique reposant sur le principe du gradient boosting. Contrairement aux approches basées sur le bagging comme les RSF, le GBSA est un modèle d'optimisation par boosting appliquée aux données censurées, qui construit les modèles de manière séquentielle, chaque modèle corrigeant les erreurs du précédent (Friedman, 2001). Dans le contexte de l'analyse de survie, le GBSA est généralement formulé comme une extension du modèle de Cox proportional hazards model, en optimisant une fonction de perte adaptée aux données censurées (Hothorn et al., 2006). Cette approche permet d'intégrer explicitement le temps jusqu'à l'événement tout en tenant compte de la censure (Pölsterl, 2015). Le processus d'estimation du risque avec le GBSA repose sur les principes du gradient boosting introduits par (Friedman, 2001) et (Bühlmann & Hothorn, 2007).

- **Deep neural network model for survival analysis (DeepSurv)**

DeepSurv est un modèle de réseau de neurones à propagation avant (feed-forward) inspiré du modèle des hasards proportionnels de Cox, auquel il apporte une amélioration remarquable en permettant de capturer des relations non linéaires entre les variables d'entrée et les événements cliniques (Katzman et al., 2018). Il généralise le modèle de Faraggi et Simon (Faraggi & Simon, 1995) en remplaçant sa structure peu profonde par un réseau de neurones feed-forward profond (FFNN), intégrant plusieurs couches cachées et des fonctions d'activation non linéaires. Le modèle est optimisé à l'aide d'algorithmes de descente de gradient et de rétropropagation, afin de maximiser une fonction de perte basée sur la vraisemblance partielle du modèle de Cox, permettant de gérer les données censurées (Katzman et al., 2018). Certaines méthodes d'analyse de survie basées sur l'apprentissage profond intègrent une fonction de perte de classement (*ranking loss*) en complément d'une fonction de perte de survie classique, afin d'améliorer les performances évaluées par l'indice de concordance (*C-index*) (Wiegrebe et al., 2024). Les différents étapes de l'estimation du risque avec DeepSurv sont décrits dans l'étude de (Katzman et al., 2018).

- 1) **Propagation dans le réseau** : Les covariables sont introduites dans un réseau de neurones à propagation avant (*feed-forward*), composé de plusieurs couches cachées.
- 2) **Modélisation du risque** : Le réseau produit en sortie un score de risque individuel (log-risk), remplaçant la composante linéaire du modèle de Cox.
- 3) **Calcul de la fonction de perte** : Une fonction de perte basée sur la vraisemblance partielle de Cox est utilisée, permettant d'intégrer les données censurées dans l'apprentissage.
- 4) **Optimisation du modèle** : Les paramètres du réseau (poids) sont ajustés via la descente de gradient pour minimiser la fonction de perte.
- 5) **Régularisations** : des techniques comme : dropout et la pénalisation L2 sont utilisées pour limiter le surapprentissage
- 6) **Estimation du risque et de la survie (deux différentes sorties)** : un score de risque individuel ou une fonction de survie, obtenu après estimation de la baseline hazard.

1.7 REVUE DE LA LITTÉRATURE

1.7.1 Stratégie de recherche documentaire

Une recherche systématique a été effectuée dans la base de données PubMed afin de garantir un processus d'analyse bibliographique rigoureux et transparent.

Cette recherche visait à identifier les études utilisant des modèles d'AA appliqués à la prédiction du risque cardiovasculaire à partir de données de survie (temps jusqu'à l'événement). Les requêtes ont combiné les termes MeSH (Medical Subject Headings) à l'aide des mots-clés suivants : cardiovascular disease*, risk*, prediction*, Survival / Time-to-Event Analysis* et Survival Machine Learning Models*.

L'équation de recherche finale a été construite à l'aide d'opérateurs booléens (AND, OR, NOT) et de combinaisons progressives. Cette approche structurée a permis d'obtenir un aperçu de l'état actuel de la recherche dans ce domaine. Grâce à la méthodologie employée, 46 études ont été initialement recensés, dont 24 répondaient aux critères d'inclusion. Les articles analysés démontrent les progrès réalisés en AA pour la prédiction des MCV. La période couverte était de 2015–2025 afin de capturer l'évolution récente des modèles d'AA de survie. Ceci est illustré dans Tableau 1 et Tableau 2.

Tableau 1. *Stratégie de recherche documentaire (PubMed, 2015–2025)*

Stratégie de recherche documentaire (PubMed de 2015-2025) – pour la revue de la littérature sur l'utilisation de l'apprentissage automatique (AA) appliquées aux données de survie pour la prédiction du risque cardiovasculaire			
Lignes	Concepts	Requêtes	Résultats
#1	Cardiovascular Diseases	("Cardiovascular Diseases"[MeSH Terms] OR "Stroke"[MeSH Terms] OR "Myocardial Infarction"[MeSH Terms] OR "Myocardial Ischemia"[MeSH Terms] OR "Heart Failure"[MeSH Terms] OR "Coronary Disease"[MeSH Terms] OR "Coronary Artery Disease"[MeSH Terms] OR "cvd"[Title/Abstract] OR "cardiovascular"[Title/Abstract] OR "cardiovascular disease*" [Title/Abstract] OR "chd"[Title/Abstract] OR "coronary heart"[Title/Abstract] OR "coronary heart disease*" [Title/Abstract] OR "cad"[Title/Abstract] OR "coronary arter*" [Title/Abstract] OR "coronary disease*" [Title/Abstract] OR "coronary artery disease*" [Title/Abstract] OR "Stroke"[Title/Abstract] OR "myocardial infarction*" [Title/Abstract] OR "myocardial ischemia*" [Title/Abstract] OR "ischemic heart"[Title/Abstract] OR "ischemic heart disease*" [Title/Abstract] OR "heart disease*" [Title/Abstract] OR "heart failure*" [Title/Abstract])	1 285 861
#2	Survival / Time-to-Event Analysis	("Survival"[MeSH Terms] OR "Survival Analysis"[MeSH Terms] OR "surviv*" [Title/Abstract] OR "survival analys*" [Title/Abstract])	915 919
#3	Prediction	("Prediction Algorithms"[MeSH Terms] OR "Risk"[MeSH Terms] OR "predict*" [Title/Abstract] OR "Risk" [Title/Abstract] OR "algorithm*" [Title/Abstract] OR "validat*" [Title/Abstract] OR "calibr*" [Title/Abstract] OR "recalibr*" [Title/Abstract] OR "discriminat*" [Title/Abstract] OR "method*" [Title/Abstract])	7 321 014
#4	Risk	("Risk"[MeSH Terms] OR "Risk Assessment"[MeSH Terms] OR "Risk Factors"[MeSH Terms] OR "risk prediction" [Title/Abstract] OR "risk model" [Title/Abstract] OR "risk models" [Title/Abstract] OR "risk assessment" [Title/Abstract] OR "risk estimation" [Title/Abstract] OR "risk stratification" [Title/Abstract] OR "risk factor*" [Title/Abstract] OR "predict*" [Title/Abstract])	2,215,861

Tableau 2. *Stratégie de recherche documentaire (PubMed, 2015–2025) (Suite)*

Stratégie de recherche documentaire (PubMed de 2015-2025) – pour la revue de la littérature sur l'utilisation de l'apprentissage automatique (AA) appliquées aux données de survie pour la prédiction du risque cardiovasculaire			
Lignes	Concepts	Requêtes	Résultats
#5	Survival Machine Learning Models	("random survival forest"[Title/Abstract] OR "gradient boosting survival"[Title/Abstract] OR "GBM-Cox"[Title/Abstract] OR "DeepSurv"[Title/Abstract] OR "DeepHit"[Title/Abstract] OR "Deep Survival Machines"[Title/Abstract] OR "Deep Survival"[Title/Abstract] OR "Deep Learning Survival"[Title/Abstract])	1,255
#6		#1 AND #2 AND #3 AND #4 AND #5	159
#7	Exclusion (diseases)	("Neoplasms"[Mesh] OR "Mental Disorders"[Mesh] OR "Depression"[Mesh] OR "Anxiety"[Mesh] OR "Dementia"[Mesh] OR "Child"[Mesh] OR "Adolescent"[Mesh] OR "Infections"[Mesh] OR "COVID-19"[Mesh] OR "Coronavirus Infections"[Mesh] OR "SARS-CoV-2"[Mesh]) OR "cancer*" [Title/Abstract] OR "tumor*" [Title/Abstract] OR "neoplasm*" [Title/Abstract] OR "oncolog*" [Title/Abstract] OR "covid" [Title/Abstract] OR "sars-cov*" [Title/Abstract] OR "infection*" [Title/Abstract] OR "infectious" [Title/Abstract] OR "virus*" [Title/Abstract] OR "bacteria*" [Title/Abstract] OR "psychiatr*" [Title/Abstract] OR "mental" [Title/Abstract] OR "depression" [Title/Abstract] OR "anxiety" [Title/Abstract] OR "alzheimer" [Title/Abstract] OR "dementia" [Title/Abstract] OR "neurodegenerative" [Title/Abstract] OR "pediatr*" [Title/Abstract] OR "child*" [Title/Abstract] OR "adolescent*" [Title/Abstract])	6 071 636
#8	Excusion mortality and hospitalization	("mortality"[Title/Abstract] OR "death"[Title/Abstract] OR "hospitalization"[Title/Abstract] OR "hospitalisation"[Title/Abstract] OR "Mortality"[MeSH Terms] OR "Hospitalization"[MeSH Terms])	1 278 885
#9	ML survival studies for cardiovascular	#6 NOT #7 NOT #8	46

1.7.2 Études antérieures

Une recension des écrits a permis d'identifier des études récentes évaluant l'apport des algorithmes d'AA de survie pour la prédiction du risque de MCV. Ces travaux comparent généralement des modèles comme RSF, DeepSurv et GBSA, à des approches statistiques traditionnelles comme la régression de Cox (CPH) ou les scores de risque établis (FRS, PCE). L'analyse de ces études révèle plusieurs tendances clés, mais aussi des lacunes méthodologiques importantes qui justifient notre projet. Afin de structurer l'analyse, les études ont été classées selon une taxonomie basée sur différents types : la taille de la population, les issues (outcomes) de survie, les différents modèles et les métriques d'évaluation. Cette approche permet une comparaison plus claire et met en évidence les limites méthodologiques de la littérature. Pour améliorer la lisibilité et de distinguer les dimensions cliniques et méthodologiques, les études ont été synthétisées à l'aide tableaux. Les tableaux 3, 4, 5, 6, 8, 9, 10 et 11 décrivent les populations, les sources de données et les issues de survie étudiée tandis que les tableaux 12, 13 et 14 détaillent les choix méthodologiques et les performances des modèles.

Performance de discrimination des modèles

Les études antérieures montrent que les modèles d'AA de survie présentent globalement une capacité de discrimination supérieure à celle des modèles traditionnels. (Ambale-Venkatesh et al., 2017; Bauer et al., 2023; Qian et al., 2023) rapportent des C-index plus élevés pour le RSF (0,80 ; 0,74 ; 0,78, respectivement) comparativement à la régression de Cox (CPH) ou au FRS. Dans les études de grandes cohortes, (Liu et al., 2025) atteignent un C-index de 0.736 avec DeepSurv, un résultat similaire à celui de (Sabovčik et al., 2022) avec GBSA, qui obtient un C-index de 0.735. Il est à noter que ces performances sont obtenues malgré des méthodologies variables et parfois incomplètes : Liu et al., 2025) n'appliquent ni sélection de variables ni l'optimisation des hyper paramètres, tandis que (Dolezalova et al., 2021), en intégrant ces étapes, rapportent une amélioration modeste, atteignant un C-index de 0.745. À l'inverse, (Teshale, Htun, Vered, Owen, Ryan, et al., 2024)

obtiennent des performances plus faibles avec RSF (C-index = 0.641 ; IBS = 0.048) et DeepSurv (C-index = 0.662 ; IBS = 0.046), et ce malgré l'utilisation d'une optimisation des hyperparamètres par Random GridSearch combinée à une validation croisée.

En revanche, sur des cohortes plus restreintes mais avec une méthodologie plus complète, les résultats sont parfois bien meilleurs. (Zhuang et al., 2022), en appliquant une sélection de variables avec la technique LASSO (Least Absolute Shrinkage and Selections Operator) et une validation croisée de 5 plis, rapportent un C-index plus élevé allant de 0.79 à 0.82 avec le modèle RSF. Ces résultats sont soutenus par (Segar et al., 2019), qui obtiennent un C-index de 0.77 en utilisant une approche similaire incluant la sélection de variables et la validation croisée. De même, (Hathaway et al., 2021), en intégrant à la fois une optimisation des hyper paramètres combinées avec une validation croisée et une sélection de variables, rapportent un AUC de 0.82 pour DeepSurv et 0.80 pour RSF. Enfin, mêmes certaines petites cohortes comme celle de (Ren et al., 2022) atteignent des C-index comparables pour DeepSurv et RSF (0.767 et 0.753), lorsque des techniques de sélection de variables sont utilisées. Cette observation renforce clairement l'idée que la simple application d'un modèle d'AA n'est pas une garantie de succès et que la qualité de la performance prédictive ainsi que sa robustesse est intrinsèquement liée à la complétude de la méthodologie appliquée.

L'hétérogénéité des performances : le cas du sexe

Compte tenu des différences biologiques en santé cardiovasculaire entre les hommes et les femmes, certaines études ont évalué les performances prédictives en fonction du sexe. Les résultats observés suggèrent des différences. Des études rapportent une capacité discriminative nettement supérieure chez les femmes (DeepSurv, C-index = 0,726 vs 0,624 chez les hommes; (Barbieri et al., 2022; Teshale, Htun, Vered, Owen, Ryan, et al., 2024): C-index = 0,81 vs 0,77). Inversement, d'autres travaux rapportent de meilleures performances chez les hommes ((Blanchard et al., 2022): AUC = 0,821 vs 0,779 ; C-index = 0,792 vs 0,742 ; (Qian et al., 2023): C-index = 0,780 vs 0,737). Ces divergences soulignent que le développement d'un modèle unique pour une population hétérogène peut masquer des écarts de performances importantes. Elles renforcent la nécessité non seulement d'évaluer

systématiquement les performances par sexe, et invite à explorer l'opportunité de développer des modèles distincts pour les femmes et les hommes, comme le propose la présente étude.

Calibration et analyse par catégorie de risque

Malgré des performances de discrimination souvent prometteuses, la quasi-totalité des études examinées présente une importante limite méthodologique : une évaluation limitée, voire absente, de la calibration des modèles et une absence d'analyse selon la catégorie de risque cardiovasculaire. Or, la calibration permet de vérifier la concordance entre le risque prédit et le risque observé, garantissant ainsi la pertinence clinique du modèle. De plus, une évaluation par catégorie de risque est essentielle pour déterminer si le modèle conserve sa performance chez les individus à risque faible, intermédiaire ou élevé. La plupart rapporte le C-index ou l'AUC, des métriques qui mesurent la capacité d'un modèle à ordonner les individus selon leur risque, mais n'informent pas sur la calibration. Cette lacune réduit la portée et la fiabilité des modèles proposés, malgré l'utilisation de cohortes importantes et de techniques d'AA avancées.

Synthèse critique

Ce survol de la revue de littérature permet de conclure que si le potentiel de l'AA de survie en prédiction du risque cardiovasculaire est indéniable, la littérature actuelle souffre d'un manque de rigueur dans l'évaluation complète des modèles et la complétude des méthodologies utilisées.

Premièrement, l'évaluation des modèles demeure souvent incomplète, se concentrant principalement sur la discrimination au détriment de la calibration et de l'utilité clinique des prédictions. Deuxièmement, les méthodologies utilisées varient considérablement d'une étude à l'autre, notamment en ce qui concerne les stratégies de validation et les métriques de performance, ce qui complique la comparaison des résultats et limite la reproductibilité des travaux. Troisièmement, peu d'études intègrent des analyses stratifiées, notamment selon le sexe ou les catégories de risque, malgré leur importance dans la compréhension des performances des modèles au sein de sous-groupes cliniquement pertinents.

Premièrement, les cohortes se répartissent en trois catégories. Les petites cohortes ($< 1\,000$ participants ; Tableau 3) proviennent surtout de contextes cliniques spécialisés et offrent une caractérisation détaillée des patients, mais avec une généralisation limitée. Les cohortes de taille moyenne (1000–9000 participants ; Tableau 4) constituent un compromis entre diversité et représentativité, malgré une hétérogénéité pouvant influencer les performances. Les grandes cohortes ($\geq 10\,000$ participants ; Tableau 5), issues de bio banques et de bases nationales, offrent une forte puissance statistique et une meilleure représentativité, permettant une validation plus robuste des modèles.

Deuxièmement, les issues cardiovasculaires suivent un continuum de complexité. Les issues composites (MCV ; Tableau 6 et ASCVD ; Tableau 9) estiment le risque global à partir de facteurs classiques, tandis que les issues spécifiques (AVC, Tableau 7 ; insuffisance cardiaque, Tableau 10 ; maladies coronariennes, Tableau 11) reposent sur des variables plus ciblées, incluant biomarqueurs et paramètres cliniques avancés. Les MACE (Tableau 8) occupent une position intermédiaire, combinant événements aigus et chroniques avec des profils de variables plus hétérogènes.

Troisièmement, les Tableaux 12 à 14 montrent une évolution vers des modèles de survie plus sophistiqués intégrant sélection de variables, optimisation et validation avancée. Les modèles d'apprentissage automatique, notamment RSF et DeepSurv, surpassent globalement les modèles classiques, mais leurs performances dépendent fortement de la qualité du pipeline méthodologique. De plus, notre analyse nous permet de croire que la qualité de la performance prédictive est intrinsèquement liée à la mise en œuvre d'une méthodologie complète, incluant des étapes comme la sélection de variables, l'optimisation des hyperparamètres, et la prise en compte de la nature déséquilibrée des données.

Tableau 3. *Études utilisant des cohortes de petite taille pour la prédiction du risque cardiovasculaire*

Références	Cohorte / source de données	Taille (N)	Population	Âge	Sexe (N / %)
Yan et al., 2025	Etude du centre unique (Second Affiliated Hospital of Anhui Medical University)	N = 318	Patients sous dialyse péritonéale	Non précisé	Non précisé
Wang et al., 2025	Etude de l'hôpital thoracique de Tianjin	N = 1202	Patients avec un infarctus du myocarde	Tranche d'âge : H ≤ 50 ans, F ≤ 55 ans	Non précisé
Chhoa et al., 2023	Cohorte CADASIL French National Referral Center for rare cerebrovascular diseases in France	N = 482	Patients atteints de CADASIL	Tranche d'âge : ≥18 ans	Non précisé
Ren et al., 2022	Étude rétrospective Patients atteints de maladie rénale diabétique (DKD)	N = 890	Patients avec maladie rénale diabétique	Non précisé	Non précisé

Tableau 4. Études utilisant des cohortes de taille moyenne pour la prédiction du risque cardiovasculaire

Références	Cohorte / source de données	Taille (N)	Population	Âge	Sexe (N / %)
Lusk et al.,2025	Données de Nashville Biosciences Lp(a) Vanderbilt University Health System	N = 4369	Patients avec un taux mesuré de Lp(a)	Âge moyen : 51 ans	F : 49.5 %
Bauer et al., 2023	Coronary CT Angiography (CAD suspects)	N = 5 457	Patients avec suspicion de coronaropathie	Âge moyen : 61 ans	H = 3648
Qian et al., 2023	Cohorte multiethnique Population rurale Xinjiang (Chine)	N = 7 975	Population Adultes sans antécédents MCV	Tranche d'âge : 30–74 ans	H = 4054 F = 3920
Morris et al., 2023	Cohorte longitudinal Jackson Heart Study The University of Mississippi Medical Center, Jackson State University, and Tougaloo College.	N = 5 300	Afro-Américains sans antécédents cardiovasculaires	Tranche d'âge : 21-94 ans	F : 64 %
Blanchard et al., 2022	Cohorte prospective retrospective (Respiratory Health Research Institute)	N = 5 506	Patients avec apnée du sommeil	Tranche d'âge : 49-69 ans	F = 39.4 %
Hathaway et al., 2021	Étude prospective de cohorte multicentrique MESA (Étude multiethnique)	N = 6814	Participants sains Multiethnique	Âge moyen : 62.52 ans Tranche d'âge : 45–84 ans	H = 46.71 %
Ishii et al., 2023	Étude de FIRE trial (<i>Atrial Fibrillation and Ischaemic Events With Rivaroxaban in Patients With Stable Coronary Artery Disease</i>)	N = 2215	Patients atteints de fibrillation auriculaire et maladie coronarienne stable	Tranche d'âge : 69–80 Âge moyen : 75 ans	H : 78.6 %
Segar et al., 2019	Cohorte Rétrospective Validation interne (ACCORD) Validation externe (ALLHAT)	N = 8 756 (ACCORD) N = 10 819 (ALLHAT)	Patients diabétiques	ACCORD : âge moyen : 62.7 ALLHAT : âge moyen : 67	ACCORD : F = 38,5 % ALLHAT : F = 49.0 %

Tableau 5. *Études utilisant des cohortes de grande taille dans la prédiction du risque cardiovasculaire*

Références	Cohorte / source de données	Taille (N)	Population	Age	Sexe (N / %)
Liu et al., 2025	Cohorte rétrospective Validation interne: cohorte UK Biobank (UKB) Validation externe: cohorte chinoise (CHERRY)	UK Biobank N = 347 206 CHERRY N = 177 756	Population sans antécédent de MCV	Tranche d'âge : 40 à 74 ans Age moyen : UKB : 58 ans, CHERRY : 55 ans	H: UKB: 45.3 % CHERRY: 46.4
Teshale et al., 2024	Cohorte prospective ASPREE (sous-échantillon)	N = 18 548	Population saine (Sans antécédents MCV)	Tranche d'âge : ≥70 ans Âge moyen = 75,4 ans	H : 8 122 F : 10 426
Deng et al., 2023	Lifetime Risk Pooling Project (6 cohortes)	N = 23 216	Population multicentrique (Noirs et blancs)	Âge moyen : 57,8	Non précisé
Barbieri et al., 2022	Bases de données nationales de santé New-Zélandaises	N = 2 164 872	Population générale	Tranche d'âge : 30–74 ans	F : 49 %
Sabovčik et al., 2022	Cohorte multi-centre (6 cohortes)	N = 30 354	Population adultes saines	Tranche d'âge : 30–80 ans Âge moyen : 66 ans	H : 66.57 %
Matheson et al., 2022	Large cohorte japonaise de 40 ans et plus	N = 155 108	Adultes japonais sains	Moyen d'âge : 47 ans Tranche d'âge : 41–56 ans	H : 65 %
Farhadian et al., 2021	Cohorte rétrospective Centre médical Farchshian, Iran	N = 503 842	Patients après une intervention coronarienne percutanée	Moyen d'âge : 47 ans Tranche d'âge : 41–56 ans	F : 31.4% H : 68.6%
Dolezalova et al., 2021	Cohorte UK Biobank (UKB)	N = 466 052	Participants sains	Âge moyen : 57 ans Tranche d'âge : 50–63 ans	H : 205379
Chun et al., 2021	Cohorte prospective China Kadoorie Biobank	N = 503 842	Population générale chinoise sans antécédents AVC	Tranche d'âge : 35–74 ans	H = 205 293 F = 298 549
Sung et al., 2019	Cohorte nationale rétrospective de la Corée du sud	N = 361 239	Population sans antécédents cardiovasculaires	Tranche d'âge : 40–79 ans	

Tableau 6. *Études utilisant les maladies cardiovasculaires globales (MCV) comme issue de survie*

Références	Variables incluses	Issue de survie
Liu et al., 2025	Âge, sexe, pression artérielle systolique, diabète, cholestérol total et tabagisme	MCV
Teshale et al., 2024	Âge, sexe, prise de médicaments antihypertenseurs, pression artérielle systolique, diabète, cholestérol à lipoprotéines de haute densité (HDL-c), non-HDL-c, créatinine sérique et tabagisme	MCV
Morris et al., 2023	Âge, sexe, profession, niveau d'éducation, tabagisme, statut diabète, IMC, LDL, HDL, cholestérol total, pression artérielle systolique et diastolique, statut assurance + facteurs psychosociaux, socio-économiques, environnementaux	MCV
Zhuang et al., 2022	Âge, sexe, profession, niveau d'éducation, tabagisme, statut diabète, IMC, LDL, HDL, cholestérol total, pression artérielle systolique et diastolique, antécédents de MCV, variables cliniques + biomarqueurs, antécédents familiaux, antécédents médicaux	MCV
Barbieri et al., 2022	Âge, sexe, ethnie, niveau de désavantage socio-économique (niveau de privation), diabète, antécédent d'hospitalisation pour fibrillation auriculaire, médicaments préventifs à la base (antihypertenseurs, hypolipémies, anticoagulants/antiplaquettaires).	MCV
Ren et al., 2022	Âge, lipoprotéines de haute densité (HDL), hémoglobine (Hb), pression artérielle systolique (PAS), tabagisme, excrétion urinaire de protéines sur 24 heures et cholestérol total	MCV
Dolezalova et al., 2021	Âge, sexe, diplôme universitaire, taille-hanche, pression artérielle systolique, cholestérol total, rapport cholestérol, médicaments, symptômes, antécédents familiaux de maladie cardiaque, tabagisme, consommation d'alcool	MCV
Sung et al., 2019	Age, sexe, pression artérielle systolique et diastolique, alcool, Cholestérol Total, tabagisme, IMC, antécédents familiaux (diabète, MCV, hypertension)	MCV

Tableau 7. *Études utilisant des Accidents Vasculaire Cérébraux (AVC) comme issue de survie*

Références	Variables incluses	Issue de survie
Yan et al., 2025	Démographiques (âge, sexe, IMC...), antécédents de diabète ou MCV, laboratoire (plaquettes, créatinine, albumine, etc.), paramètres écho cardiographiques	AVC
Lusk et al., 2025	Âge, sexe, tension artérielle, tabagisme, taux de lipides, antécédents médicaux, antécédents MCV, dépression, antihypertenseur, race, ethnie, consommation alcool, diabète	AVC
Chhoa et al., 2023	Âge, sexe, antécédents de maladies, HDL, LDL, cholestérol total, tabagisme, niveau d'éducation, consommation alcool, pression artérielle systolique ou diastolique, pression artérielle systolique et diastolique, trouble cognitif, trouble psychiatrique, autres antécédents médicaux	AVC
Ishii et al., 2023	Âge, sexe, indice de masse corporelle, pression artérielle systolique, consommation d'alcool, clairance de la créatinine, insuffisance cardiaque, diabète, utilisation d'antiagrégants et type de FA	AVC
Barbieri et al., 2022	Âge, sexe, ethnie, niveau de désavantage socio-économique (niveau de privation), diabète, antécédent d'hospitalisation pour fibrillation auriculaire, médicaments préventifs à la base (antihypertenseurs, hypolipémies, anticoagulants/antiplaquettaires).	AVC
Chun et al., 2021	Facteurs sociodémographiques, alimentation, les antécédents médicaux, activité physique et des mesures cliniques.	AVC

Tableau 8. *Études utilisant des événements cardiovasculaires majeurs indésirables (MACE) comme issue de survie*

Références	Variables incluses	Issue de survie
Wang et al., 2025	Âge, sexe, indice de masse corporelle, hypertension, diabète, hyperlipidémie et maladie coronaire antérieure, médicaments, inhibiteurs de l'ECA, cholestérol total, LDL, HDL, triglycérides, créatinine et HbA1c, tabagisme, consommation alcool, douleur thoracique, dyspnée, paramètres électrocardiographiques, antécédents familiaux MCV	MACE
Bauer et al., 2023	Âge, sexe, antécédents de maladies, HDL, LDL, cholestérol total, tabagisme, hypertension, antécédents familiaux, IMC	MACE
Blanchard et al., 2022	Signaux physiologiques de sommeil + variables cliniques	MACE
Farhadian et al., 2021	Âge, diabète, tabagisme, longueur de stent, maladie coronarienne aiguë ou chronique et hyperlipidémie	MACE

Tableau 9. *Études utilisant événements cardiovasculaires athérosclérotiques (ASCVD) comme issue de survie*

Références	Variables incluses	Issue de survie
Deng et al., 2023	Diabète, traitement HTA, tabac, HDL, cholestérol total, âge, sexe	ASCVD
Qian et al., 2023	Âge, poids, taille, statut matrimoniale, tabagisme, consommation alcool, hypertension, diabète, LDL, HDL, cholestérol total, profession, niveau d'éducation	ASCVD
Matheson et al., 2022	Antécédents de maladie cardiaque, âge, prise antihypertenseurs, cholestérol HDL, sexe, antécédents de maladies, HDL, LDL, cholestérol total, tabagisme, IMC, race, pression artérielle systolique et diastolique, artérielle, pression artérielle systolique et diastolique, autres antécédents médicaux	ASCVD
Hathaway et al., 2021	Age, sexe, race, diabète, pression artérielle systolique et diastolique, consommation alcool, Cholestérol Total, tabagisme, IMC, antécédents familiaux, niveau d'éducation, revenu, médication hypertension, diabète	ASCVD

Tableau 10. *Études utilisant l'Insuffisance cardiaque (IC) comme issue de survie*

Références	Variables incluses	Issue de survie
Yan et al., 2025	Démographiques (âge, sexe, IMC...), antécédents de diabète ou MCV, laboratoire (plaquettes, créatinine, albumine, etc.), paramètres écho cardiographiques	Insuffisance cardiaque,
Sabovčik et al., 2022	Variables cliniques, démographiques, biométriques	Insuffisance cardiaque
Barbieri et al., 2022	Âge, sexe, ethnie, niveau de désavantage socio-économique (niveau de privation), diabète, antécédent d'hospitalisation pour fibrillation auriculaire, médicaments préventifs à la base (antihypertenseurs, hypolipémies, anticoagulants/antiplaquettaires).	Insuffisance cardiaque
Segar et al., 2019	Age, sexe, antécédents de maladies, HDL, LDL, cholestérol total, tabagisme, IMC, race, pression artérielle systolique ou diastolique, artérielle, pression artérielle systolique et diastolique, autres antécédents médicaux, antécédents MCV... etc.	Insuffisance cardiaque
	Age, sexe, race, diabète, pression artérielle systolique et diastolique, consommation alcool, Cholestérol Total, tabagisme, IMC, antécédents familiaux, niveau d'éducation, revenu, médication hypertension, diabète	Insuffisance cardiaque,

Tableau 11. *Études utilisant les maladies coronariennes comme issue de survie*

Références	Variables incluses	Issue de survie
Lusk et al.,2025	Âge, sexe, tension artérielle, tabagisme, taux de lipides, antécédents médicaux, antécédents MCV, dépression, antihypertenseur, race, ethnie, consommation alcool, diabète	Maladies coronarienne
Zhuang et al., 2022	Âge, sexe, profession, niveau d'éducation, tabagisme, statut diabète, IMC, LDL, HDL, cholestérol total, pression artérielle systolique et diastolique, antécédents de MCV, variables cliniques + biomarqueurs, antécédents familiaux, antécédents médicaux	Maladies coronariennes
Matheson et al., 2022	Antécédents de maladie cardiaque, âge, prise antihypertenseurs, cholestérol HDL, sexe, antécédents de maladies, HDL, LDL, cholestérol total, tabagisme, IMC, race, pression artérielle systolique et diastolique, artérielle, pression artérielle systolique et diastolique, autres antécédents médicaux	Maladies coronariennes

Tableau 12. Synthèse des approches méthodologiques et performances des modèles
(Quelques études utilisant des techniques de sélection et ingénierie de variables)

Références	Modèle(s) AA de survie	Comparateurs/ Baseline	Sélection et ingénierie de variables	Métriques d'évaluation	Performances
Yan et al., 2025	RSF	XGBoost	LASSO	AUC C-index	RSF: C-index = 0,725 XGBoost: C-index = 0.703
J.-X. Wang et al., 2025	RSF DeepSurv	CPH XGBoost	RSF LASSO	C-index Brier score (BS)	RSF: C-index=0.815, BS=0.125
Chhoa et al., 2023	RSF GBSA	CPH Elastic Net Cox	LASSO, RFE Unvariable Cox	AUC BS	RSF : AUC = 0.773 , BS = 0.172 GBSA : AUC = 0.766, BS = 0.165 CPH : AUC = 0.770, BS : 0.173
Qian et al., 2023	RSF	CPH FRS	L1 regularized Cox regression, RSF	AUC C-index BS	RSF H: AUC = 0.791, C-index = 0.780, BS = 0.060 F: AUC = 0.76, C-index = 0.737, BS = 0.110
Ishii et al., 2023	RSF	CPH FRS	Algorithme Boruta + pondération par HR	AUC BS	RSF AUC = 0,70, C-index = 0.706, BS = 0.80
Hathaway et al., 2021	DeepSurv RSF	SVM CPH	Oui (Tehcnique non précisé)	AUC	DeepSurv : AUC=0.82 RSF : AUC=0.80 CPH : AUC=0.80
Dolezalova et al., 2021	DeepSurv	CPH	Tree-Structured Parzen Estimator (TPE)	C-index	DeepSurv: C-index = 0,7446 , H: C-index =0.72, F: C-index = 0.75 CPH: C-index = 0,7443
Chun et al., 2021	RSF	SVM, Gradient Boosted Trees (GBT), CPH MLP	Oui (Technique non précisée)	AUC	H : AUC (GBT : 0.833, RSF : 0.826, SVM : 0.83) F : AUC (GBT : 0.836, RSF : 0.832, SVM : 0.831)
Ambale- Venkatesh et al., 2017	RSF	CPH, AIC-Cox, FRS, ACC/AHA ASCVD	RSF Forward Selection Backwards Selection	C-index BS	MCV globaux RSF : C-index = 0.80, BS = 0.079 CPH : C-index = 0.76, BS = 0.087 ACC/AHA ASCVD RSF : C-index = 0.73, BS = 0.089 FRS : C-index = 0.73, BS = 0.11

Tableau 13. Synthèse des approches méthodologiques et performances des modèles
(Quelques études utilisant l'optimisation des hyperparamètres + la validation croisée)

Références	Modèle(s) AA de survie	Comparateurs/ Baseline	Optimisation des hyperparamètres	Validation	Métriques d'évaluation	Performances
Teshale et al., 2025	RSF DeepSurv	CPH	Random GridSearch	Validation croisée K = 5 plis	C-index BS	RSF (C-index = 0,641, BS = 0,048) DeepSurv (C-index = 0,662, BS= 0,046)
Bauer et al., 2023	RSF	CPH FRS	Recherche par grille (GridSearch)	Nested cross-validation (CV) Interne : K= 4plis Externe : K= 4plis	C-index	RSF = 0.74 CPH = 0,71 FRS = 0.65
Qian et al., 2023	RSF	CPH FRS	Oui (Technique non précisée)	Validation croisée (K non précisé)	AUC C-index BS	RSF: H : AUC=0.791, C-index = 0.780, BS = 0.060 F : AUC = 0.76, C-index = 0.737, BS = 0.110
Chhoa et al., 2023	RSF GBSA	CPH Elastic Net Cox	Oui (Technique non précisée)	Nested cross-validation (CV) Interne: K= 4 plis Externe: K= 10 plis	AUC BS	RSF : AUC = 0.773, BS = 0.172 CWGB : AUC = 0.766, BS = 0.165 CPH : AUC = 0.770, BS = 0.173
Deng et al., 2023	DeepSurv Nnet-survival	Pooled Cohort Equations (PCE) Cox PH-TWI	Oui (Technique non précisée)	Nested cross-validation (CV) Interne: K= 10 plis Externe: K= 10 plis	C-index	Validation interne : DeepSurv H : 0,7371 ; F : 0,7972 Validation externe H : 0,7032 ; F : 0,7972
Chun et al., 2021	RSF	support vector machines (SVM) Gradient Boosted Trees (GBT), CPH, MLP (perceptrons multicouches)	Oui (Technique non précisée)	Validation croisée KFold (K = 5 plis)	AUC	H : AUC (GBT : 0.833, RSF : 0.826 , SVM : 0.83) F : AUC (GBT : 0.836, RSF : 0.832 , SVM : 0.831)

Tableau 14. Synthèse des approches méthodologiques et performances des modèles (Quelques études sans l'optimisation des hyperparamètres et sans sélection et ingénierie des variables)

Références	Modèle(s) AA de survie	Compareurs/ Baseline	Sélection et ingénierie de variables	Optimisation hyper paramètres	Métriques d'évaluation	Performances
Liu et al., 2025	DeepSurv DeepHit	CPH	Non précisé	Non précisé	C-index, BS	C-index (DeepSurv : 0.733 CPH : 0.731)
Lusk et al., 2025	RSF	CPH	Non précisé	Non précisé	C-index	RSF : 0.82 CPH: 0.69
Blanchard et al., 2022	DeepSurv	CNN	Non précisé	Non précisé	AUC C-index	H : AUC = 0.821 C-index = 0.792 F : C-index= 0.742, AUC = 0.779
Farhadian et al., 2021	RSF	CPH	Non précisé	Non précisé	C-index BS	RSF: IBS = 0,124, C-index = 0,648 CPH :C-index=0,626, IBS = 0,135
Sung et al., 2019	RSF DeepSurv	CPH XGBoost	Non précisé	Non précisé	AUC, BS	RNN-LSTM survival F : AUC = 0.79 H : AUC = 0.75

CHAPITRE 2 : SURVIVAL MACHINE LEARNING FOR 10-YEAR CARDIOVASCULAR RISK PREDICTION USING FRAMINGHAM SCORE SIMPLIFIED : A CANADIAN COHORT STUDY

2.1 RÉSUMÉ EN FRANÇAIS DE L'ARTICLE

La prédiction du risque de maladies cardiovasculaires (MCV) constitue un enjeu majeur de santé publique, en raison de leur forte prévalence et de leur impact sur la morbidité et la mortalité à long terme. Les modèles statistiques traditionnels, tels que le score de Framingham (FRS), bien que largement utilisés en pratique clinique, présentent certaines limites, notamment dans leur capacité à modéliser des interactions complexes entre les facteurs de risque et le risque cardiovasculaire. L'objectif de cette étude est de développer et de comparer les performances prédictives de modèles d'AA de survie pour la prédiction du risque cardiovasculaire à 10 ans, en s'appuyant sur les prédicteurs du score de Framingham simplifié (FRS-SP). Dans notre méthodologie, nous comparons trois modèles de survie : le Random Survival Forest (RSF), le Gradient Boosting Survival Analysis (GBSA) et le DeepSurv. Nous intégrons l'optimisation des hyperparamètres et le sous-échantillonnage de la classe majoritaire, pour traiter la nature déséquilibrée des données, garantissant ainsi une représentation plus équilibrée des différentes classes. Cette approche aide à optimiser les performances du modèle, à réduire les biais et à renforcer la robustesse. De plus, l'optimisation des hyperparamètres est réalisée à l'aide d'Optuna, garantissant que chaque algorithme fonctionne à son niveau optimal. Les performances des modèles ont été évaluées à l'aide de l'indice de concordance (C-index), du score de Brier, des courbes de calibration ainsi que des courbes AUC-ROC pour l'évaluation de la discrimination. Les résultats indiquent que les modèles RSF et GBSA présentent les meilleures performances, tant dans la population globale que lors des analyses stratifiées par sexe. Toutefois, leurs performances demeurent globalement meilleures à celles des outils de prédiction traditionnels existants, tout en offrant une meilleure flexibilité pour la modélisation de interactions complexes entre les facteurs de risque. Notre article a été finalisé et sera soumis dans la revue JMIR plus précisément dans **JMIR Cardio** accessible en open source.

Survival Machine Learning for 10-Year Cardiovascular Risk Prediction Using Framingham-Simplified Predictors: A Canadian Cohort Study

Amy LEYE,¹ Mehdi Adda PhD,¹ Mahée Gilbert-Ouimet, PhD,² Xavier Trudel, PhD,³

¹Department of Mathematics, Computer Science, and Engineering, University of Quebec at Rimouski, 300, Ursulines Lane, Rimouski, QC G5L 3A1, Québec, Canada

² Department of Health Sciences, University of Quebec at Rimouski (UQAR), Rimouski (QC), Canada

³ Department of Social and Preventive Medicine, Faculty of Medicine, University Laval (QC), Canada

Corresponding Author:

Amy LEYE

04-209, Rue des Pins Est, Québec (QC), Canada G1L1T1

Phone: (581) 246-9797

Email: leya0002@uqar.ca

Abstract

Background: Cardiovascular disease (CVD) remains the leading cause of death globally. Cardiovascular risk prediction is an essential tool for guiding treatment and prevention. Several cardiovascular risk prediction tools, including the Framingham Risk Score (FRS), were developed and validated. A simplified version of the FRS, the Framingham Risk Score Simplified (FRS-SP), replaces lipids with body mass index and was recently validated in the Canadian population, demonstrating acceptable predictive performance. Like most available risk prediction tools based on traditional survival models, the FRS-SP fails to capture complex non-linear relationships and interactions among predictors. Survival Machine Learning (ML) models may help overcome these limitations. However, studies examining the performance of survival ML models using the same set of clinical variables as traditional survival models remain limited.

Objective: Using a rigorous reproducible framework, we develop and compare three survival ML models (RSF, GBSA, and DeepSurv) trained on the same set of clinical predictors as the FRS-SP (age, sex, systolic blood pressure, BMI, diabetes status, antihypertensive medication use, and smoking status), and evaluate their calibration and discrimination for 10-year CVD risk prediction.

Methods: This study analyzed 6,883 participants (49.5% women, mean age 47.0) from the prospective PROQ cohort, using FRS-SP predictors measured at baseline (1999–2001). Three survival ML models (RSF, GBSA, and DeepSurv) were trained using 5-fold cross-validation with rigorous hyperparameter optimization (Optuna) and class imbalance handling. Performance was evaluated via C-index, Brier score, calibration plots and stratified by sex.

Results: RSF and GBSA outperformed DeepSurv benefiting from hyperparameter tuning for data imbalance. Overall, both models demonstrated strong discrimination (C-index = 0.73, AUC = 0.74) and excellent calibration with low Brier scores (0.0779 and 0.0785, respectively), and mean calibration values close to 1 (1.02 and 1.03, respectively). In contrast, DeepSurv demonstrated lower discrimination (C-index = 0.67) and poor calibration, with evidence of risk overestimation (mean calibration = 1.71 overall). This poor calibration is particularly significant in women (mean calibration = 2.99), which is directly relevant to stratified analyses. Furthermore, the

analysis by risk category reveals a massive overestimation in the high-risk category across all subgroups, while an underestimation is observed in women in the intermediate-risk category. Stratified analyses by sex showed strong performance in both women and men, with excellent mean calibration (RSF = 1.01 and GBSA = 1.005).

Conclusions: Ensemble-based survival ML models (RSF and GBSA) achieved acceptable discrimination and excellent calibration for 10-year CVD risk prediction using FRS-SP predictors in a Canadian prospective cohort, whereas DeepSurv substantially miscalibrated risk, particularly among women.

To our knowledge, this is the first study to apply survival ML (RSF, GBSA, and DeepSurv) using the predictor set of an established clinical risk score — the FRS-SP — within a Canadian prospective cohort, and to directly compare their calibration performance in this setting using a fully reproducible methodological framework (Bayesian hyperparameter optimization, stratified cross-validation, calibration-focused evaluation, sex-stratified performance).

Keywords: Cardiovascular disease; Risk Prediction; Time-to-Event Analysis; Survival Machine Learning; Random survival forests; Deep learning survival; Gradient boosting survival

Introduction

Cardiovascular disease (CVD) remains the leading cause of death and disability worldwide, with nearly 19.8 million deaths in 2022, accounting for 32% of all deaths [1,2]. In Canada, approximately 8.5% of adults were affected by CVD in 2017–2018 [3]. Accurate cardiovascular risk assessment is essential to identify individuals most susceptible to developing CVD and to guide prevention strategies [4]. Canadian guidelines (C-CHANGE 2022) on primary prevention in cardiology recommend the use of risk prediction models to assess cardiovascular risk and guide clinical practice [5]. Available and well-known models include SCORE [6], PCE [7], QRISK [8], PREVENT [9] and the Framingham Risk Score (FRS) [10]. Developed within the Framingham Heart Study [11], the FRS estimates the 10-year risk of CVD based on recognized risk factors [12,13]. Within the original validation study, a simplified version substituting BMI (Body Mass Index) for lipids (FRS-SP) has shown performance comparable to the conventional model [14]. Despite its potential utility in both clinical and nonclinical settings, the FRS-SP model shares some of the limitations of the FRS, including a tendency to overestimate risk in women [15] and a limited ability to capture nonlinear or interactions between risk factors [16]. However, findings from the Canadian validation study by Garçon et al. [17] suggest an opposite tendency, with an underestimation of risk in women, particularly among those at high risk. These conflicting observations suggest that miscalibration in women may vary across populations and modeling approaches, highlighting the need for prediction methods that better capture sex-specific risk patterns.

Models based on machine learning (ML) emerge as a promising tool to improve CVD risk prediction. They allow analyzing large volumes of data, modeling nonlinear interactions [18–21] and identifying the most relevant risk predictors [22]. Available studies generally report a higher discriminatory ability of ML models compared to traditional statistical approaches such as Cox regression [23,24] or established risk scores [25]. To date, three survival ML models; RSF, DeepSurv, and GBSA, were thoroughly evaluated in previous studies and have demonstrated strong predictive performance [26–29]. Despite an increasing interest in ML to predict

cardiovascular risk, the literature presents several important methodological limitations. First, performance evaluation is usually incomplete, with most studies focusing on discrimination measures (e.g., C-index, AUC) while ignoring calibration [30,31] and performance assessment by risk category, which are essential for clinical decisions. Secondly, heterogeneity is observed in model development and evaluation procedures, with the highest performance generally achieved when models are based on a reproducible methodology including variable selection, hyper parameter optimization, and cross-validation [32–34]. Studies that omit some of these steps report lower performance [35–37] and greater variability in performance; moreover, this limits comparability between studies. Thirdly, comparisons by sex yield inconsistent findings: some studies report better predictive performance in women [38–40], whereas others report better performance in men [41,42], suggesting that population-level prediction approaches may mask differences in predictive accuracy between biological sexes. The present study aims to address these gaps by proposing a comprehensive and transparent methodological framework for the prediction of CVD based on machine learning.

This study develops and compares the predictive performance of survival ML models (RSF, GBSA, and DeepSurv) for 10-year cardiovascular risk prediction in a prospective cohort of 6,883 men and women. The secondary objectives were to assess the model's discrimination and calibration, as well as to analyze its performance by sex and defined risk categories (low: < 10%, intermediate: 10–20%, high: > 20%).

Methods

Study Design and Population

This study utilized data from the PROspective Quebec (PROQ) Study on Work and Health, a prospective cohort described in detail elsewhere [43]. This cohort included, at baseline, 9,188 white-collar employees (aged 18–65 years) recruited from 19 public organizations in the Quebec region between 1991 and 1993 (T1; 75% participation rate). Follow-up data were collected at T2 (1999–2001; 89% participation) and T3 (2015–2018; 81% participation). The mean follow-up duration was 24 years from T1 to T3. Linkage to medico-administrative databases was achieved for 96% of all participants.

The current analysis focused on the cardiovascular component of the PROQ cohort present at the 1999–2001 data collection (T2), which was defined as the study baseline. Participants were excluded if they had a history of CVD at baseline, those aged outside the 30–74 range. The final analytical sample consisted of 6,883 participants (mean age: 47.0 years; 49.5% women) followed over a 10-year period.

Baseline Characteristics

Predictors were selected based on the Simplified Framingham Risk Score (FRS-SP) for primary care [14]. Data were obtained via self-administered questionnaires and physical measurements at T2:

- **Demographics:** Age (continuous) and Sex (categorical).
- **Clinical Measurements:** Systolic Blood Pressure (SBP) was measured at rest using a sphygmomanometer following American Heart Association guidelines [44]. The average of two initial measurements was used. Body Mass Index (BMI) was calculated (kg/m^2) and measured by trained personnel.
- **Comorbidities & Lifestyle-related factors:** Diabetes status (Yes/No), Antihypertensive medication use (Yes/No), and Current Smoking status (Yes/No).
- **Risk Stratification:** Participants were categorized into low (<10%), moderate (10–19%), and high ($\geq 20\%$) risk groups according to Canadian Cardiovascular Society thresholds [45].

Outcome Definition: Cardiovascular Disease

The primary outcome was the 10-year incidence of a first cardiovascular event, defined as Ischemic Heart Disease (IHD), Stroke/Transient Ischemic Attack (TIA), Congestive Heart Failure (CHF), or Peripheral Artery Disease (PAD). Events were identified in Quebec's medico-administrative databases (RAMQ, MED-ÉCHO) using validated algorithms [46–50]. Diagnostic codes were extracted using ICD-9 and ICD-10 standards, and procedure codes via CCI and CCP classifications. Detailed case definitions are provided in **Supplementary Table S1**.

Model development

We developed three survival models based on machine learning and compared them to the FRS-SP predictors:

1. **RSF**: An ensemble method extending random forests to right-censored data using log-rank splitting rules [51,52].
2. **GBSA**: A non-parametric boosting ensemble optimizing a survival loss function (Cox partial likelihood) [53,54].
3. **DeepSurv**: A feed-forward deep neural network (Cox-based loss) designed to model non-linear interactions [55,56].

For each model, the predicted risk was calculated from its estimated cumulative hazard function, and the observed 10-year risk was obtained using the cumulative incidence function. The models were implemented in Python (version 3.11.7, 64-bit) using scikit-survival, PySurvival, lifelines, and PyTorch.

The proposed methodological approach for predicting cardiovascular risk is structured in five (5) steps, as illustrated in Figure 1. **First step**: Data were collected from 6,833 participants aged 30 to 74, then cleaned and preprocessed by the CHUL team, including the handling of missing values, the coding of categorical variables, and the selection of variables (risk factors). **Second step**: The dataset was subsequently divided into two subsets: a training/validation set (70%) and a test set (30%). **Third step**: Model development was performed on the training set, using 5-fold cross-validation combined with hyperparameter optimization via Optuna. Given the class imbalance, analyses were conducted on both the original data and on data rebalanced

through random undersampling of the majority class. **Fourth step:** Three ML models for survival analysis (RSF, GBSA, and DeepSurv) were trained on the overall population as well as stratified by sex (women and men). **Fifth step:** The models performance was evaluated using different metrics, including discrimination (C-index, AUC), calibration, and the Brier score. Individuals were also classified into risk categories ($< 10\%$, $10\text{--}19\%$, $\geq 20\%$).

The work presented in this article was conducted in accordance with the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines [57].

Experimental Protocol

Data Splitting: The dataset was partitioned into a training/validation set (70%) and a held-out test set (30%). To prevent data leakage, all preprocessing (imputation, scaling) was fitted solely on the training set and applied to the test set.

Hyperparameter Optimization: Optuna, a Bayesian optimization framework [58], was combined with 5-fold cross-validation on the training set to identify optimal hyperparameters (See Supplementary Table S3).

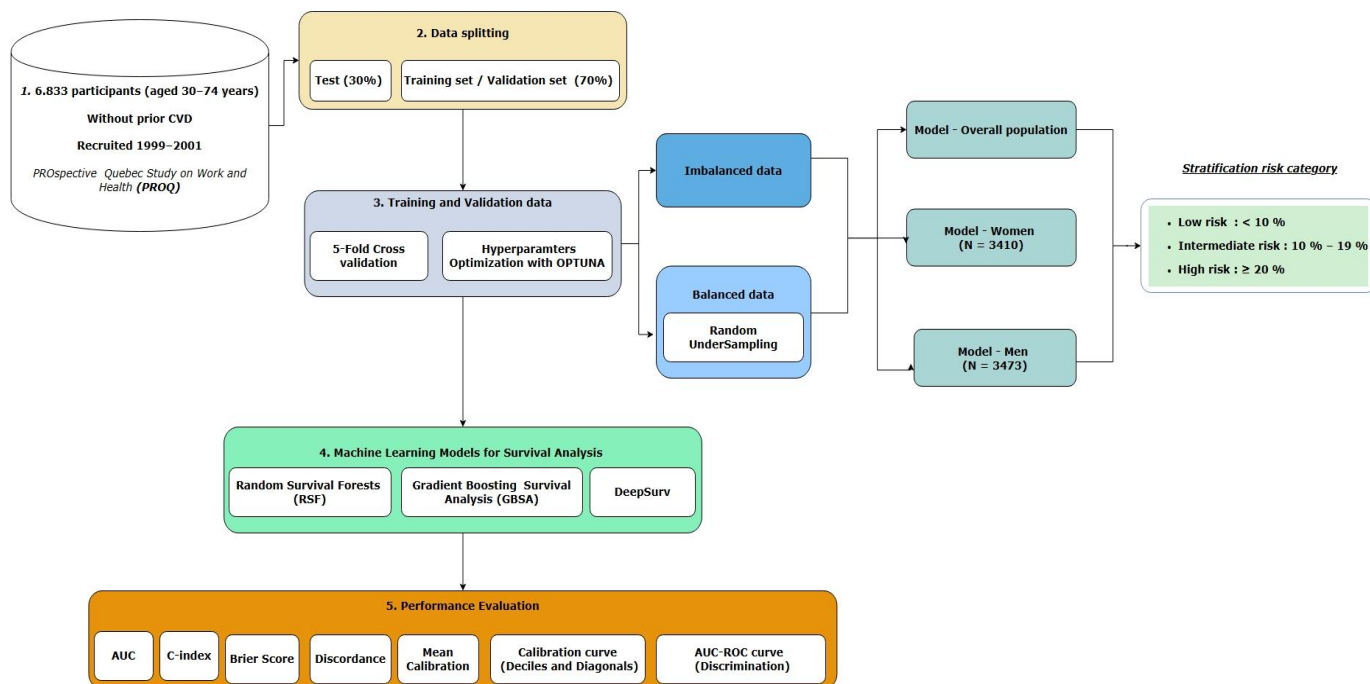
Handling Class Imbalance: A significant class imbalance was observed, with 90.9% non-MCV and 9.1% MCV cases (see Supplementary Figure S1). This imbalance may bias the models toward the majority class, reducing performance on the minority class [59]. To mitigate the impact of this imbalance, two distinct strategies were implemented:

1. **Hyperparameter Tuning:** Hyperparameters were tuned to improve minority class detection performance, including `class_weight`, `min_samples_leaf`, and `n_estimators` (tree density) [60,61].
2. **Random Undersampling:** Random undersampling was applied to the majority class within the training folds to balance class distribution [62]. Both strategies were evaluated on the independent test set.

Performance evaluation

Calibration and discrimination were used, as recommended by Ben Van Calster et al [63] and Steyerberg et al [64] to evaluate the performance of prediction models. Calibration was assessed using mean calibration and discordance [65]. Calibration was also examined qualitatively using calibration curves (decile and diagonal approaches). Discrimination was assessed using two performance measures: the C-index (C-statistics) [66,67] and the area under the ROC (Receiver Operating Characteristic) curve, widely used to evaluate the discriminatory capacity of a model [68]. Overall performance was quantified by the Brier score, which simultaneously integrates calibration and discrimination and whose low values reflect greater accuracy [64]. Finally, stratified analyses were also performed according to sex and risk categories (<10%, 10-19%, $\geq 20\%$) to assess the robustness of performance within different subgroups.

Figure 1. Data analysis flowchart



Ethics considerations

All participants provided informed consent, and the ethics committee of the CHU de Québec – University Laval Research Center, approved the study. Data can be made available by the corresponding author on reasonable request.

Results

Baseline Characteristics of the Study Population

A total of 6,883 participants were included in the analysis (Table 2), the sex distribution was balanced, with 3,473 men (50.5%) and 3,410 women (49.5%). The mean age was 47.0 years (standard deviation: 7.9), slightly higher in men (48.5 years) than in women (45.6 years). Regarding comorbidities, 2.6% of participants had diabetes, 14.3% were smokers, and 7.1% reported taking antihypertensive treatment. Smoking was more common in women (15.8%) than in men (12.8%). Mean systolic blood pressure was 120.4 mmHg (SD: 14.91), higher in men (125.2 mmHg) than in women (115.5 mmHg). During the 10-year follow-up, 626 cardiovascular events occurred (men = 426, women = 200).

Table 1. Baseline characteristics of participants (N = 6883)

Characteristics	Overall	Men	Women
	(N = 6883)	(N =3473, 50.5%)	(N= 3410, 49.5%)
Mean Age $\pm$ SD, years	47.1 $\pm$ 7.9	48.5 $\pm$ 8.3	45.6 $\pm$ 7.4
Diabetes status, n (%)			
No	6704 (97.4 %)	3372 (97.1 %)	3332 (97.7 %)
Yes	179 (2.6 %)	101 (2.9 %)	78 (2.3 %)
Smoking status, n (%)			
No	5901 (85.7 %)	3030 (87.2 %)	2871 (84.2 %)
Yes	982 (14.3 %)	443 (12.8 %)	539 (15.8 %)
Antihypertensive medication, n (%)			
No	6392 (92.9 %)	3198 (92.1 %)	3194 (93.7 %)
Yes	491 (7.1 %)	275 (7.9 %)	216 (6.3 %)
Mean SBP $\pm$ SD, mmHg	120.4 $\pm$ 14.9	125.2 $\pm$ 13.8	115.5 $\pm$ 14.3
Mean BMI $\pm$ SD, kg/m ²	25.6 $\pm$ 4.3	26.4 $\pm$ 3.7	24.9 $\pm$ 4.7
Incident CVD events, n (%)			
No	6257 (90.9 %)	3047 (87.7 %)	3210 (94.1 %)
Yes	626 (9.1 %)	426 (12.3 %)	200 (5.9 %)

Table 1 footnote. Abbreviation used: BMI: Body Mass Index; SBP: Systolic Blood Pressure; SD: Standard Deviation. Data are presented as mean $\pm$ SD for continuous variables, and number (percentage) for categorical variables.

Results of Predictive Models

Calibration and Discrimination (Overall Population)

In the overall population, including men and women, the RSF model predicted a risk of 9.27%, close to the observed risk of 9.10%, corresponding to a mean calibration of 1.02 (**Table 2**). The calibration graphs (**Figure 2A and 3A**) show that the curve closely follows the diagonal. Discrimination, assessed by the C-index, was 0.73, with a Brier score of 0.0779, which corresponds to acceptable performance according to criteria commonly used for survival prediction models. The GBSA model produced a predicted risk of 9.33%, also close to the observed risk of 9.10%. The calibration curves (**Figure 2B and 3B**) show slight differences compared to those of RSF. Discrimination, represented by the C-index and the Brier score are comparable to those of the RSF model. In contrast, the DeepSurv model overestimated risk, with a predicted value of 15.58% versus 9.10% observed. Its discrimination, assessed by the C-index (0.67), as well as the calibration curves (**Figure 2C and 3C**), indicate performance inferior to that of the RSF and GBSA models. The discriminant performance of the three models, illustrated by the areas under the curve (AUC) in Supplementary **Figure S2**, were 0.74 for RSF and GBSA, and 0.67 for DeepSurv. These results reveal that RSF and GBSA present comparable and adequate discrimination for the overall population, while DeepSurv shows slightly lower discrimination. All results of analyses performed with the random undersampling technique are represented in **Supplementary Material (Figure S5, S6, S7 and S8)**.

Calibration and Discrimination (Sex-Stratified)

In men, the RSF model predicted a 10-year cardiovascular risk of 12.41%, close to the observed risk of 12.28%, corresponding to a mean calibration of 1.01. Discrimination, measured by the C-index, was 0.70 and the Brier score was 0.1025, indicating acceptable performance. The GBSA model gave a predicted risk of 12.35%, very close to the observed risk of 12.28%, with a mean calibration of 1.005, a C-index of 0.71 and a Brier score of 0.1024, values comparable to those of RSF.

In women, the RSF model predicted a risk of 5.79% versus 5.87% observed, corresponding to a mean calibration of 1.01. The C-index was 0.71 and the Brier score was 0.0536. The GBSA model predicted a risk of 5.90%, close to the observed risk of

5.87%, with a mean calibration of 1.005, a C-index of 0.72 and a Brier score of 0.0544. These results are presented in **Table 2**. The calibration graphs (**Figure 4**) show a slight underestimation of risk in women with higher risk levels for both models. The discriminant performance of the RSF, GBSA and DeepSurv models, illustrated by the areas under the curve (AUC) in **Figure 5**, was 0.70, 0.71 and 0.69 in men and 0.71, 0.73 and 0.70 in women, indicating overall comparable discrimination between RSF and GBSA, while DeepSurv presents slightly lower performance.

Table 2. Calibration and Discrimination of the models (RSF, GBSA, DeepSurv) for 10-Year Cardiovascular Diseases Prediction, Overall, Women and Men

Population	Predicted (%)	Observed (%)	Absolute difference (%)	Discordance (%)	Mean calibration	Brier score	C-index	AUC
Overall								
RSF	9.27	9.10	0.17	2	1.02	0.0779	0.73	0.74
GBSA	9.33	9.10	0.23	3	1.03	0.0785	0.73	0.74
DeepSurv	15.58	9.10	6.48	71	1.71	0.1388	0.67	0.67
Men								
RSF	12.41	12.28	0.13	1	1.01	0.1025	0.70	0.70
GBSA	12.35	12.28	0.06	0.5	1.005	0.1024	0.71	0.71
DeepSurv	20.98	12.28	8.70	71	1.71	0.1732	0.69	0.69
Women								
RSF	5.79	5.87	-0.07	1	1.01	0.0536	0.71	0.71
GBSA	5.90	5.87	0.04	0.5	1.005	0.0544	0.72	0.73
DeepSurv	17.56	5.87	11.70	199	2.99	0.1342	0.70	0.70

Table 2 footnote. The mean calibration is calculated as the ratio of predicted event rates to observed event rates. Discordance is calculated using the formula: $\text{Discordance} = [(\text{Predicted Risk} - \text{Observed Risk}) / \text{Observed Risk}] \times 100$. The Absolute difference is the difference between the predicted mean event rate and the

observed mean event rate (Predicted - Observed). RSF, random survival forest; GBSA, gradient boosting survival analysis.

Stratification by risk group (Overall Model and Sex-Stratified Model)

In the overall population (**Table 3**), the RSF model shows better overall agreement between predicted risks and observed rates, particularly in the low- and high-risk categories (e.g., 24.81% vs 23.19% in the high-risk group) in contrast, the GBSA model, exhibited a tendency to underestimate risk in the intermediate and high-risk categories (**Supplementary Figure S3**).

In women (**Figure 6**), the model RSF shows a slight underestimation in the intermediate risk category (13.19% predicted vs 12.83% observed). While the GBSA model largely overestimates risk in this same category, (12% predicted vs 24.24% observed).

In contrast, DeepSurv overestimates risk in the intermediate category (7.83% predicted vs 14.48% observed) and largely underestimates the high-risk category (65.72% predicted vs 12.64% observed).

In men (**Figure 6**), the RSF model underestimates risk in both in the high-risk category (14.39% predicted vs 15.57% observed) and the intermediate-risk category (12.60% predicted vs 15.01% observed). Similarly, the GBSA model also shows a slight underestimation in the high-risk category (23.18% predicted vs 25% observed).

Table 3. Predicted and Observed 10-Year Cardiovascular Disease Risk by Risk Category

	Low risk (<10%)			Intermediate risk (10-19%)			High risk (≥ 20%)		
Population	N(%)	predicted (%)	Observed (%)	N(%)	predicted (%)	observed(%)	N(%)	predicted (%)	Observed (%)
Overall									
RSF	1412 (68)	5.15	4.82	377(18)	13.34	14.85	276(13)	24.81	23.19
GBSA	1545(75)	7.09	5.11	394(19)	12.90	18.78	126(6)	25.72	27.78
DeepSurv	1308(63)	2.39	5.58	218(11)	14.44	11.93	539(26)	61.48	16.51
Men									
RSF	535(51)	6.62	6.54	334(32)	14.39	15.57	173(17)	26.49	23.70
GBSA	453(43)	8.37	5.30	433(42)	12.60	15.01	156(15)	23.18	25
DeepSurv	672(64)	7.43	1.78	70(7)	14.21	12.94	300(29)	65.63	23
Women									
RSF	836(82)	4.14	4.31	187(18)	13.19	12.83	0	0	0
GBSA	990(97)	5.70	5.25	33(3)	12	24.24	0	0	0
DeepSurv	702(69)	2.11	3.37	90(9)	7.83	14.48	231(22)	65.72	12.64

Table 3 footnote. Risk categories are based on the Framingham-simplified risk score, which estimates the 10-year probability of developing CVD.

The categories are defined are follows: The table presents the mean predicted risk according to the RSF, GBSA and DeepSurv and the observed event rate over a 10-year follow-up period, both for the overall population and stratified by sex (Men and Women).

Figure 2. Comparison of Predicted and Observed 10-Year CVD Risk across Deciles of Risk (Overall Population)

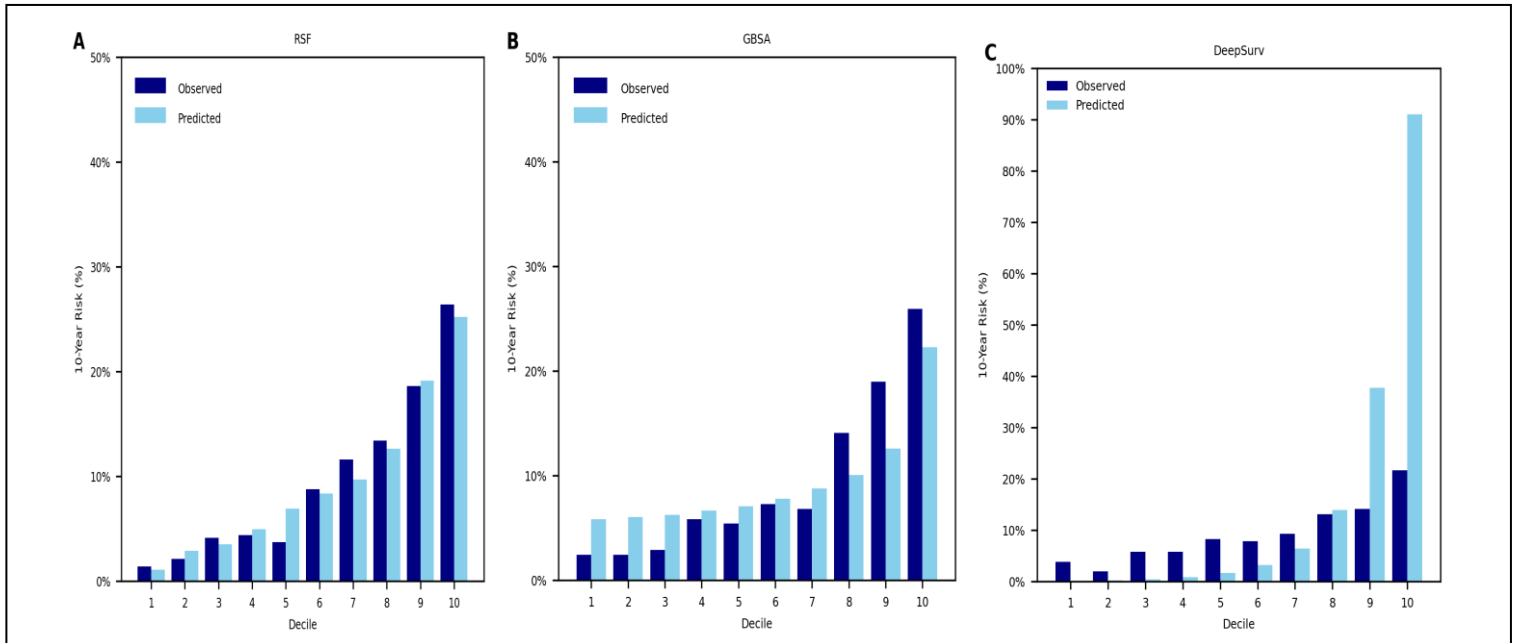


Figure 3. Calibration Plots Comparing Predicted and Observed 10-Year CVD Risk (Overall Population)

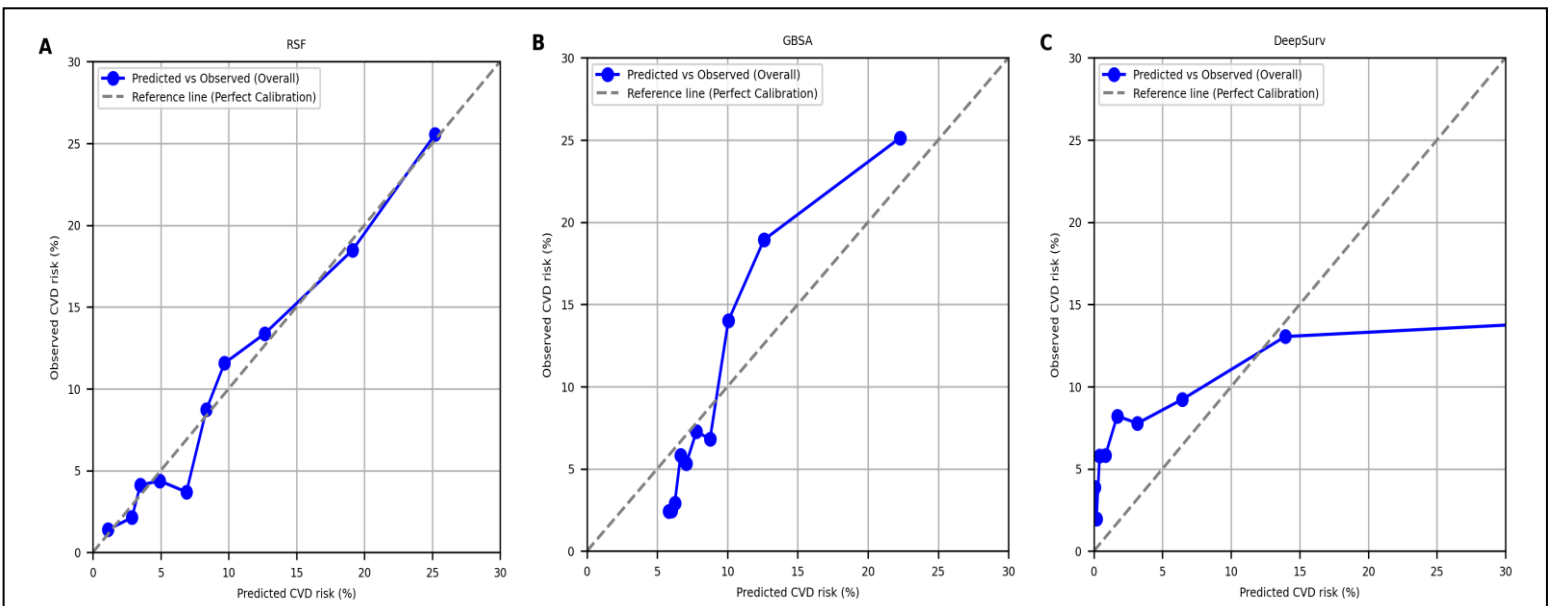


Figure 4. Comparison of Predicted and Observed 10-Year CVD Risk Across Deciles of Risk (Women and Men)

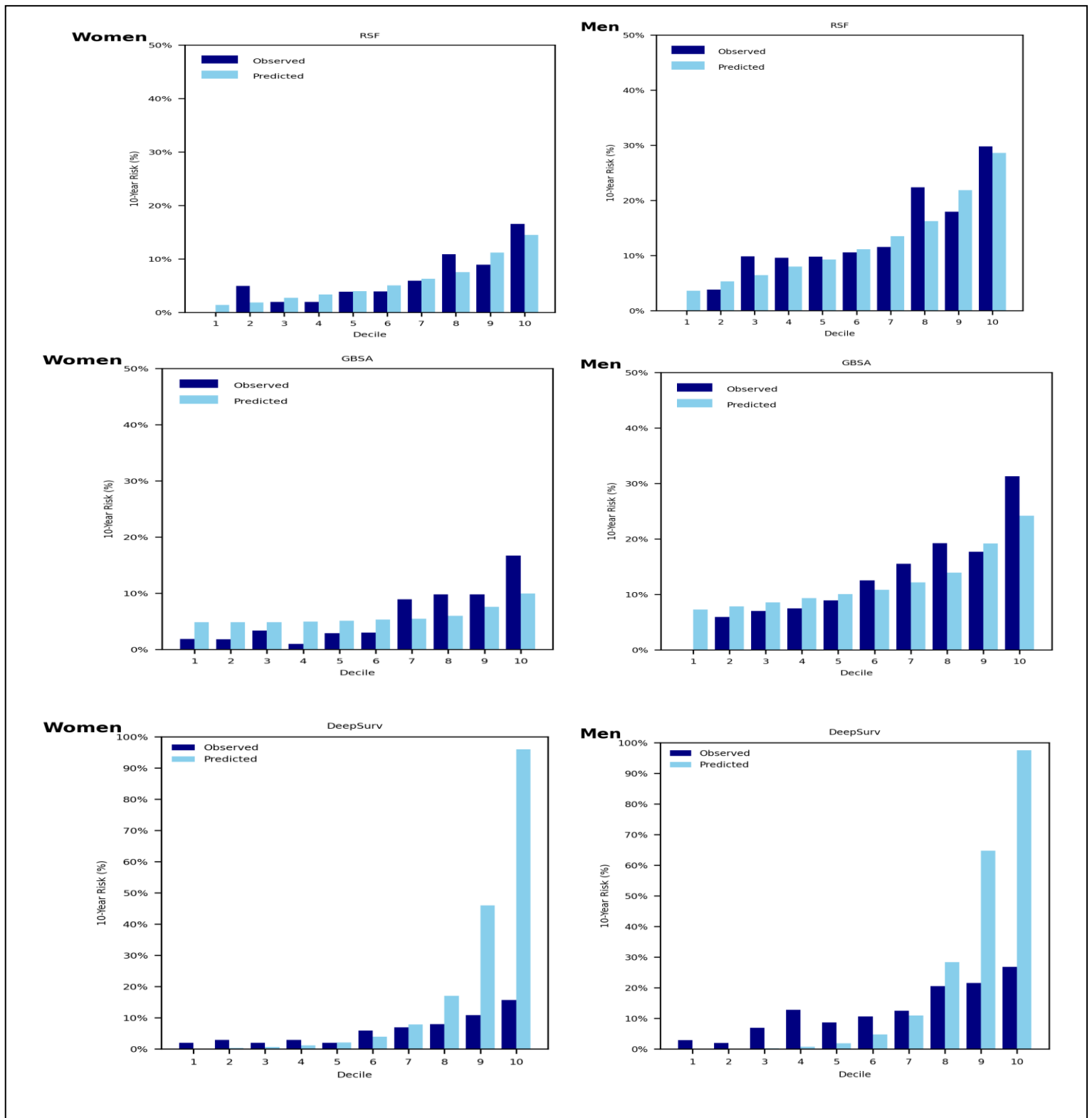
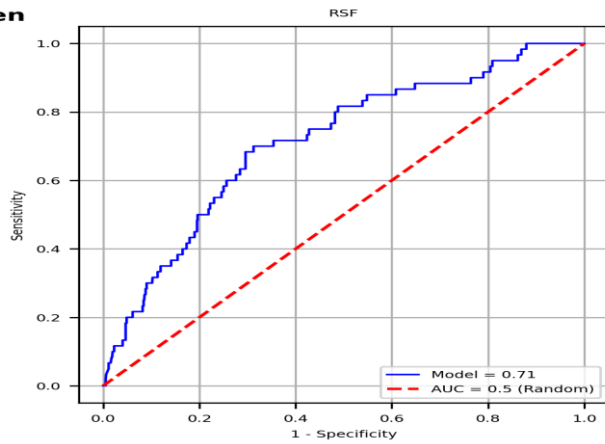
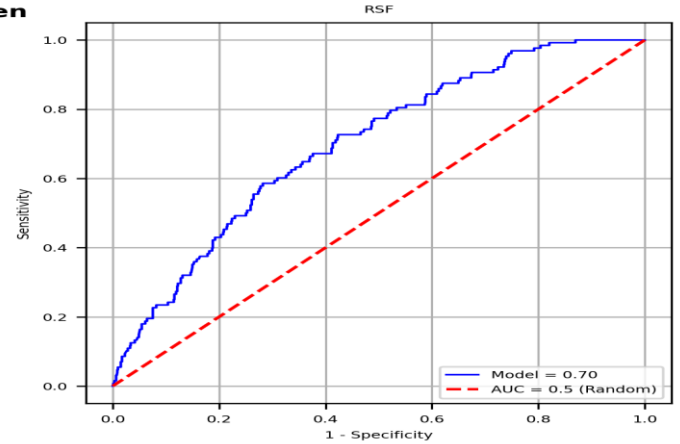


Figure 5. Receiver Operating Characteristic Curves for Cardiovascular Disease Risk Prediction Models (Women, and Men)

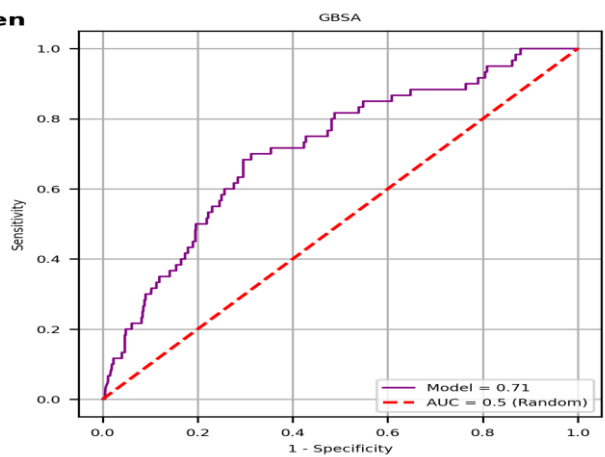
Women



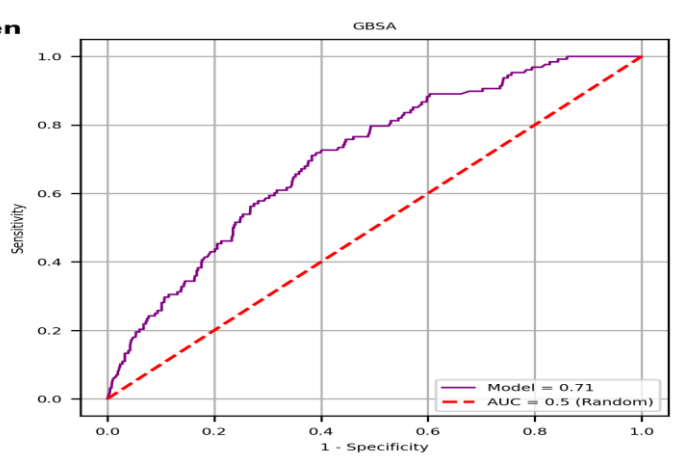
Men



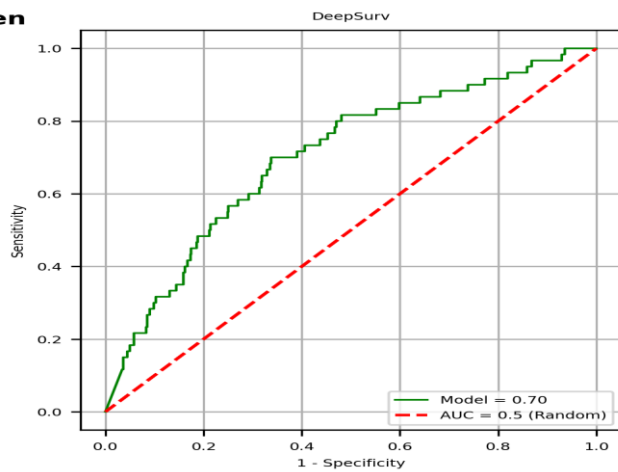
Women



Men



Women



Men

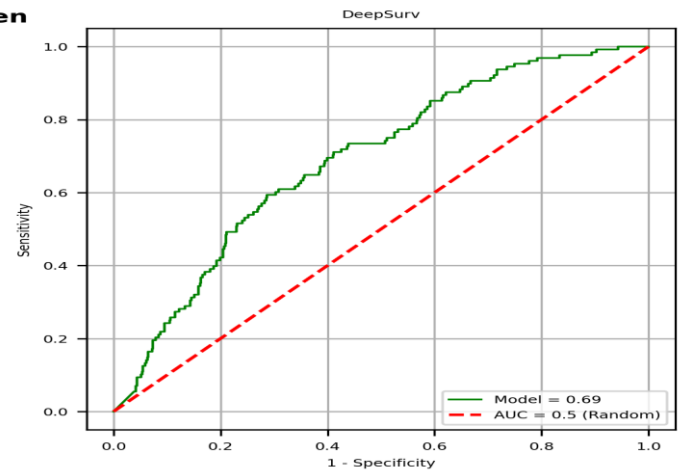


Figure 6. Predicted and Observed 10-Year Cardiovascular Disease Risk by Risk Category (Women and Men)

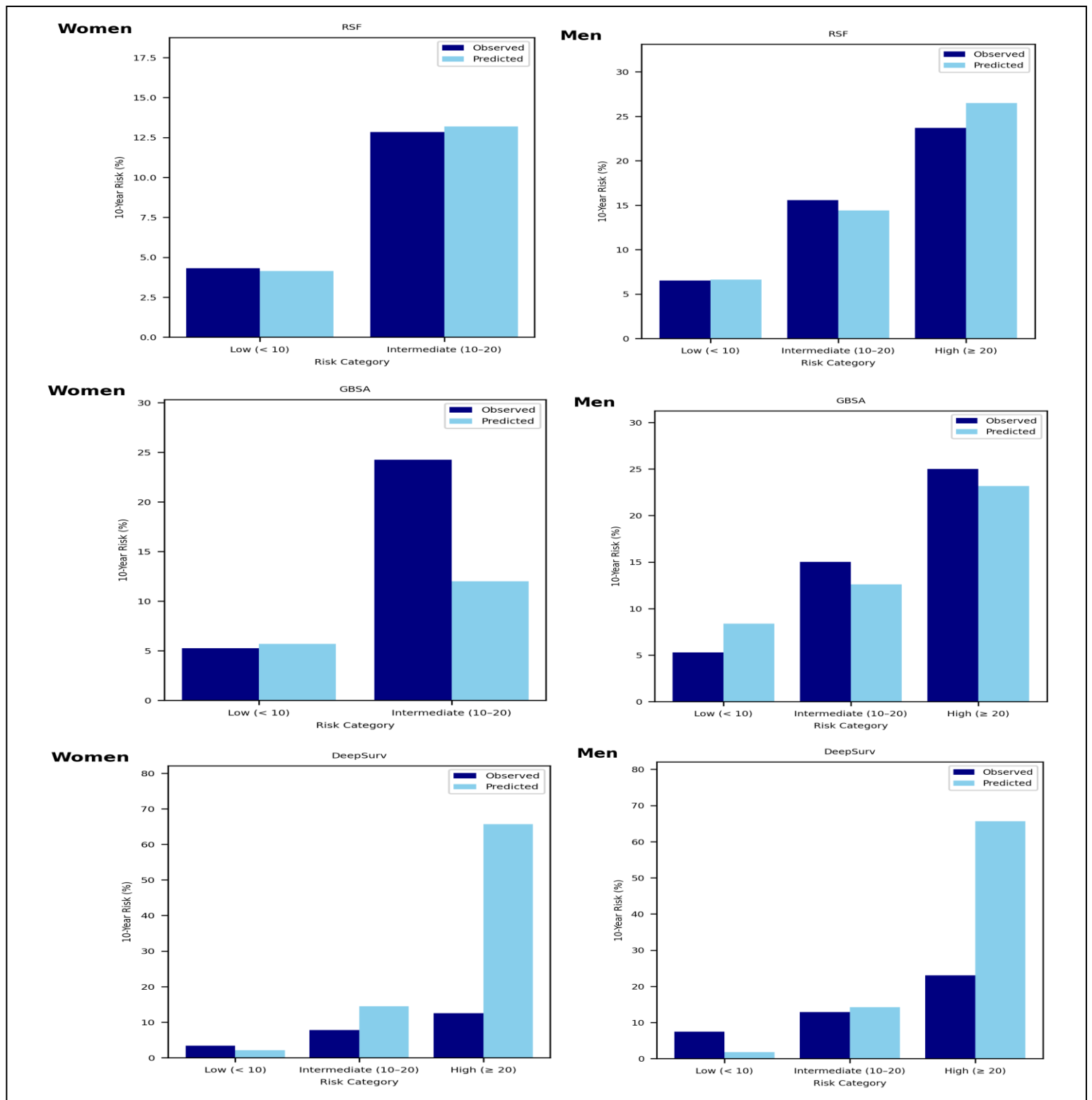
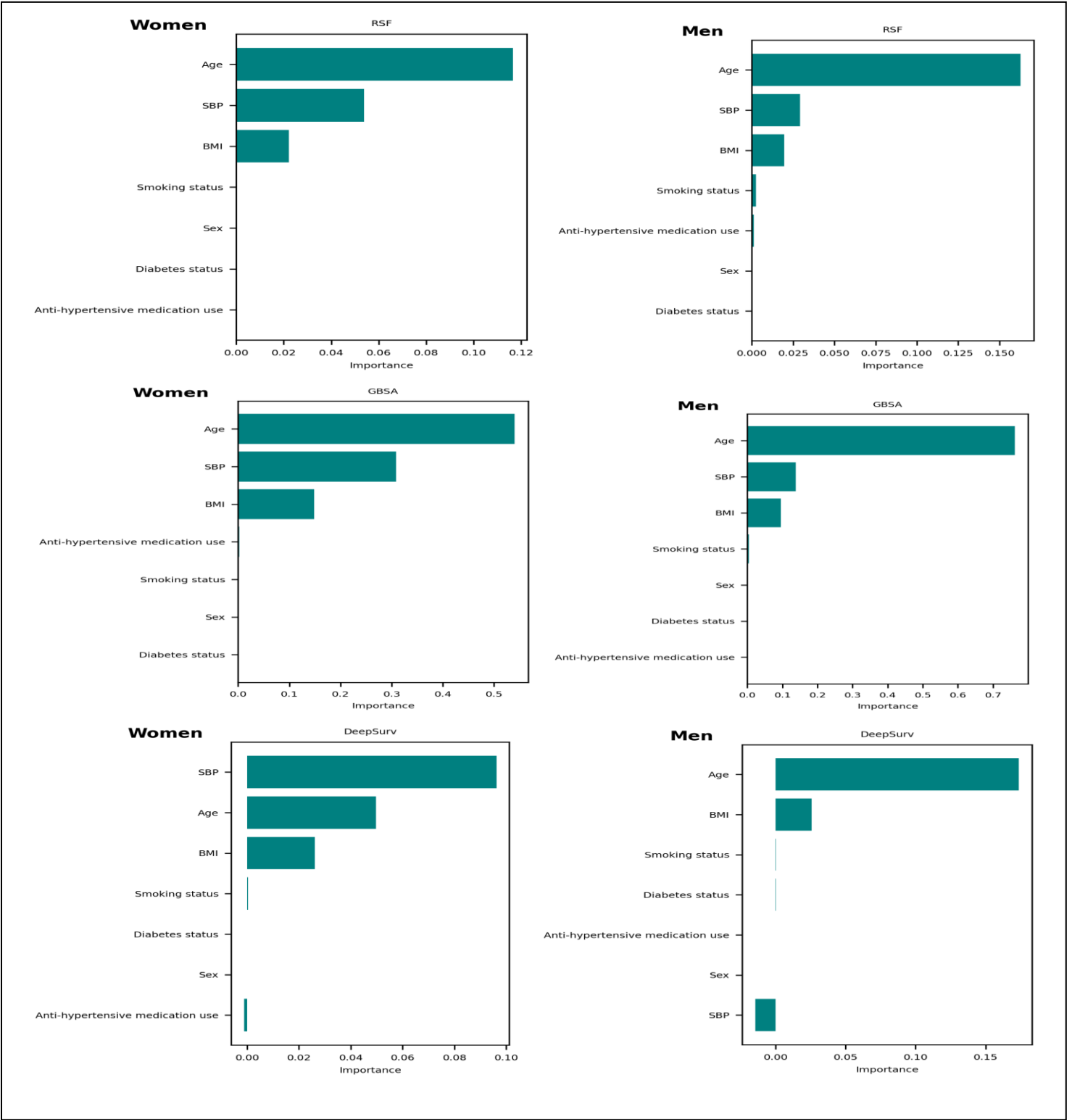


Figure 7. Comparative Analysis of Variable Contributions to Cardiovascular Risk Prediction Models (Women and Men)



Discussion

Principal Results

In this study, we evaluated the performance of survival ML models (RSF, GBSA, and DeepSurv) for predicting the 10-year cardiovascular risk in a cohort of 6,883 participants. The study aimed to assess the predictive performance of ML models in terms of calibration and discrimination using predictors from the simplified version of the Framingham tool [14]. To our knowledge, we are the first study to assess the performance of survival ML models for CVD risk prediction based on predictors derived from existing tools. The RSF and GBSA models demonstrated the best discriminative capabilities (C-index=0.73 and AUC=0.74). The mean calibration was 1.02 for RSF and 1.03 for GBSA. A low Brier score was obtained (RSF = 0.0779 and GBSA = 0.0785), which accounts for both calibration and discrimination, also indicating overall prediction accuracy consistent with that of well-calibrated clinical models [65]. We observed these performances both in the overall model as well as in women and men. In contrast, DeepSurv showed a significant overestimation of cardiovascular risk, with the overestimation reaching up to about five times the true value in certain subgroups, particularly among women.

Comparison with Previous Studies

Previous validation studies of survival ML models for cardiovascular risk prediction reported C-indices of approximately 0.74 [69] [40], consistent with the present finding. Other studies [25,28,39] have also observed high performance, supporting that survival ML models can constitute relevant alternatives for cardiovascular risk prediction. In our study, RSF showed better discrimination and calibration, contrary to some previous studies where DeepSurv obtained the best performance [33,37]. Several factors may explain the overestimation of risk observed with DeepSurv, including the limited effective sample size for deep learning models, the marked imbalance between classes in the high-risk subgroups of women, and the challenges associated with applying neural network-based survival models to relatively small, stratified clinical datasets. Our results corroborate the 'Tabular Data' hypothesis: on structured data with limited

sample size ($N < 10^4$), tree-based ensembles (RSF/GBSA) generally outperform neural networks due to better handling of discrete features and lower data volume requirement [70]. Li et al. [71] compared classical ML models (which do not account for censoring) with traditional survival models combined with predictors from the FRS on a 10-year follow-up horizon, and showed that ML models were more likely to overestimate risk, whereas survival models provided more consistent estimates by explicitly considering time to event and censoring. The fact that our study accounts for censoring therefore strengthens the robustness of our results and fully aligns with the recommendations formulated by Li et al. [71].

Two Canadian validation studies of a large multiethnic cohort [72,73] also evaluated the discrimination and calibration of the FRS. Sud et al. [72] successfully demonstrated that recalibration reduced risk overestimation with a C-index of 0.77 in women and 0.74 in men. Meanwhile, Ko et al. [73] showed risk overestimation with an overall C-index of 0.74. Both of these studies were conducted over a 5-year period. In addition, Garçon et al. [17] reported a global C-index of 0.72 for the FRS-SP model in a Canadian cohort. This study was conducted on the same population as the one used in our study. Calibration analyses showed an underestimation of risk across all risk categories, in both women and men. Compared to the study by Garçon et al., our results show a slight improvement in the discrimination performance of the ML models, indicating better risk stratification using the same predictors. Furthermore, better calibration was observed in our models. In conclusion, the integration of survival ML models applied to the FRS-SP predictors improved both discrimination and calibration in cardiovascular risk prediction.

Strengths and Limitations

Our study presents several methodological and clinical strengths. First, we directly compared three (3) ML survival models (RSF, DeepSurv, and GBSA) using the same predictors from an existing tool. Second, we applied a rigorous methodology integrating management of imbalanced data, cross-validation, and hyperparameter optimization, thereby reinforcing the robustness and reproducibility of the results. Third, we evaluated model performance not only in terms of discrimination (C-

index/AUC) but also in terms of calibration (Brier score and calibration curves), which is essential for clinical confidence. Finally, the stratified analysis by risk categories and by sex strengthens the clinical relevance of the results.

However, certain limitations must be highlighted. As a primary prevention cohort, the population was predominantly healthy at baseline, which accurately reflects the target demographic for screening tools like FRS. However, this may limit the generalizability of the models. While these models discriminate effectively at the population level, their accuracy may be compromised when applied to populations with more comorbidities. Moreover, the underrepresentation of women in the high-risk category limits the robust evaluation of calibration within this subgroup.

Additional limitations warrant mention. First, our analyses relied on internal validation only; external validation in independent cohorts is necessary before clinical adoption. Second, the PROQ cohort consists predominantly of white-collar workers from the Quebec City region, which limits generalizability to more ethnically and socioeconomically diverse populations. Third, competing risks (non-cardiovascular mortality) were handled through standard right-censoring rather than Fine-Gray models, which may marginally affect risk estimates in older participants. Fourth, we did not conduct sensitivity analyses varying key modeling choices (e.g., alternative hyperparameter ranges, different class-imbalance strategies), which could further characterize model stability.

Clinical and Public Health Implications.

The strong calibration of RSF and GBSA using only seven routinely collected predictors suggests these models could be implemented in primary care without requiring new measurements beyond those already used for the FRS-SP. Practical deployment could take the form of a web-based calculator or an integrated electronic health record tool, providing clinicians with individual 10-year CVD risk estimates alongside decile-level calibration plots to support risk communication with patients. Such tools may be particularly valuable in non-clinical settings (e.g., workplace screening programs, community pharmacies), where FRS-SP-style tools are already used. However, our findings in high-risk female subgroups highlight the importance

of subgroup-specific validation and explicit caution warnings before any tool based on these models is deployed in practice.

Future work

Several directions warrant future investigation. First, external validation in independent Canadian and international cohorts is essential before clinical deployment, as is recalibration for populations with different baseline risk. Second, explainability techniques (SHAP values, local interpretable models) should be applied to improve clinical acceptability and support shared decision-making.

Third, feature engineering incorporating psychosocial, socioeconomic, and environmental determinants may further improve predictive performance, particularly in underrepresented subgroups. Fourth, prospective implementation studies evaluating the integration of these models into primary-care workflows — including their effect on risk communication and clinical decisions — are needed to assess their real-world clinical utility.

Conclusions

In this study, we developed, validated, and compared three ML models applied to survival data to predict cardiovascular risk. These models successfully combined high discrimination with precise calibration (particularly in men), whereas the deep neural network approach (DeepSurv) showed its limitations in correctly calibrating risk. Beyond raw performance, this research highlights the critical importance of modeling strategy when facing imbalanced data. Our study establishes that fine hyperparameter optimization better preserves calibration quality than under sampling techniques, which are often counterproductive in this context. The future integration of these models into clinical practice must, however, remain vigilant regarding performance in high-risk women subgroups, suggesting avenues for improvement through cohort enrichment.

Acknowledgments

This study was funded by the CHUL Research Center at Saint-Sacrement Hospital in collaboration with the Population Health and Optimal Health Practices (SPPOS) research program. The authors also acknowledge the support of the University of Quebec at Rimouski (UQAR), where this work was conducted as part of the first author's graduate research.

Authors' Contributions

Leye Amy: Conceptualization, Methodology, Writing – review & editing.

Adda Mehdi (Director of Research): Supervision, Conceptualization, Methodology, Validation, Writing – review & editing.

Trudel Xavier: Supervision, Validation, review & editing.

Mahée Gilbert-Ouimet: Supervision, Validation, review & editing.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The PROQ cohort data used in this study are not publicly available due to ethical restrictions and privacy protections established at the time of participant consent. De-identified data may be made available from the corresponding author upon reasonable request and subject to approval by the CHU of Quebec City - Laval University Research Center Ethics Committee and a formal data-sharing agreement.

Code Availability

The Python code used to train, validate, and evaluate the survival ML models (including hyperparameter optimization scripts and calibration analyses) is publicly available on GitHub at [https://github.com/leyeamy2018/cvd-survival-analysis/blob/master/AmyLEYE Code Manuscript 20 04 2026.ipynb](https://github.com/leyeamy2018/cvd-survival-analysis/blob/master/AmyLEYE%20Code%20Manuscript%2004%202026.ipynb)

References

1. World Health Organization. Cardiovascular diseases (CVDs) [Internet]. Geneva: WHO; 2023 [cited Jan 22, 2026]. Available from: [https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
2. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. May 16, 2016 ;353:i2416.
3. Public Health Agency of Canada. *Surveillance of heart disease and other cardiac conditions* [Internet]. 2017 [cited Nov 30, 2024]. Available from: <https://www.canada.ca/en/public-health/services/diseases/heart-health/heart-disease-other-cardiac-conditions/surveillance-heart-disease-cardiac-conditions.html>
4. Lloyd-Jones DM. Cardiovascular Risk Prediction. *Circulation*. Apr 20, 2010;121(15):1768-77. doi:10.1161/CIRCULATIONAHA.109.849166
5. Jain R, Stone JA, Agarwal G, Andrade JG, Bacon SL, Bajaj HS, et al. C-CHANGE guideline for the harmonization of national cardiovascular disease prevention and management guidelines in primary care in Canada: 2022 update. *CMAJ*. Jan 2023;195(1):E21–42.
6. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J*. Jun 2003;24(11):987-1003. doi:10.1016/s0195-668x(03)00114-3 PubMed PMID: 12788299.
7. Preiss D, Kristensen SL. The new pooled cohort equations risk calculator. *Can J Cardiol*. May 2015;31(5):613-9. doi:10.1016/j.cjca.2015.02.001 PubMed PMID: 25843167.

8. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*. Jul 2007;335(7611):136. doi:10.1136/bmj.39261.471806.55 PubMed PMID: 17615182; PubMed Central PMCID: PMC1925200.
9. Khan SS, Matsushita K, Sang Y, Ballew SH, Grams ME, Surapaneni A, et al. Development and Validation of the American Heart Association's PREVENT Equations. *Circulation*. Febr 6, 2024;149(6):430-49.
10. Anderson TJ, Grégoire J, Pearson GJ, Barry AR, Couture P, Dawes M, et al. 2016 Canadian Cardiovascular Society Guidelines for the Management of Dyslipidemia for the Prevention of Cardiovascular Disease in the Adult. *Canadian Journal of Cardiology*. Nov 1, 2016;32(11):1263-82. doi:10.1016/j.cjca.2016.07.510
11. Andersson C, Naylor M, Tsao CW, Levy D, Vasan RS. Framingham Heart Study: JACC Focus Seminar, 1/8. *J Am Coll Cardiol*. Jun 2021;77(21):2680-92. doi:10.1016/j.jacc.2021.01.059 PubMed PMID: 34045026.
12. Cybulska B, Kł osiewicz-Latoszek L. Landmark studies in coronary heart disease epidemiology. The Framingham Heart Study after 70 years and the Seven Countries Study after 60 years. *Kardiol Pol*. 2019;77(2):173-80. doi:10.5603/KP.a2019.0017 PubMed PMID: 30828782.
13. Yusuf S, Hawken S, Ôunpuu S, Dans T, Avezum A, Lanas F, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *The Lancet*. Sept 11, 2004;364(9438):937-52. doi:10.1016/S0140-6736(04)17018-9 PubMed PMID: 15364185.
14. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. Feb 12, 2008;117(6):743-53. doi:10.1161/CIRCULATIONAHA.

15. Premasagar P, Aldous C, Esterhuizen TM, Gomes BJ, Gaskell JW, Tabb DL. Comparing conventional statistical models and machine learning in a small cohort of South African cardiac patients. *Informatics in Medicine Unlocked*. Jan 1, 2022;34:101103. doi:10.1016/j.imu.2022.101103
16. Bao L, Wang Y tong, Zhuang J ling, Liu A jun, Dong Y jun, Chu B, et al. Machine Learning–Based Overall Survival Prediction of Elderly Patients With Multiple Myeloma From Multicentre Real-Life Data. *Front Oncol*. Jun, 30 2022;12. doi:10.3389/fonc.2022.922039
17. Garçon J, Lavigne-Robichaud M, Milot A, Talbot D, Gilbert-Ouimet M, Brisson C, et al. Initial Cardiovascular Disease Risk Assessment in Nonclinical Settings: Framingham-Simplified Model Predictive Performance in a Canadian Cohort. *CJC Open*. Feb 19, 2026. doi:10.1016/j.cjco.2026.01.014
18. Hill P, Jonsson D, Lederman J, Bolin P, Vicente V. Uncovering nonlinear patterns in time-sensitive prehospital breathing emergencies: an exploratory machine learning study. *BMC Med Inform Decis Mak*. Jun 3, 2025;25:205. doi:10.1186/s12911-025-03046-z PubMed PMID: 40462078; PubMed Central PMCID: PMC12135243.
19. Khudhur DY, Shibghatullah AS, Shaker K, Abdul Latif A, Muda ZC. Recent Trends in Machine Learning for Healthcare Big Data Applications: Review of Velocity and Volume Challenges. *Algorithms*. Dec 2025;18(12):772. doi:10.3390/a18120772
20. Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. *Journal of Integrative Bioinformatics*. Sep 1, 2018;15(3). doi:10.1515/jib-2017-0030
21. Taha K. Machine learning in biomedical and health big data: a comprehensive survey with empirical and experimental insights. *J Big Data*. Mar 7, 2025;12(1):61. doi:10.1186/s40537-025-01108-7
22. Kapsner LA, Feißt M, Purbojo A, Prokosch HU, Ganslandt T, Dittrich S, et al. Using Machine Learning and Feature Importance to Identify Risk Factors for Mortality in Pediatric Heart Surgery. *Diagnostics (Basel)*. Nov 18, 2024;14(22):2587.

doi:10.3390/diagnostics14222587 PubMed PMID: 39594254; PubMed Central PMCID: PMC11592497.

23. Jin Y, Zhao M, Su T, Fan Y, Ouyang Z, Lv F. Comparing Random Survival Forests and Cox Regression for Nonresponders to Neoadjuvant Chemotherapy Among Patients With Breast Cancer: Multicenter Retrospective Cohort Study. *J Med Internet Res*. Apr, 18 2025;27:e69864. doi:10.2196/69864 PubMed PMID: 40198909; PubMed Central PMCID: PMC12015342.
24. Ishii M, Kaikita K, Yasuda S, Akao M, Ako J, Matoba T, et al. Risk prediction score for clinical outcome in atrial fibrillation and stable coronary artery disease. *Open Heart*. May 2023;10(1):e002292. doi:10.1136/openhrt-2023-002292 PubMed PMID: 37173099; PubMed Central PMCID: PMC10186465.
25. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular Event Prediction by Machine Learning. *Circulation Research*. Oct 13, 2017;121(9):1092-101. doi:10.1161/CIRCRESAHA.117.311312
26. Lusk JB, O'Brien EC, Hammill BG, Li F, Mac Grory B, Patel MR, et al. Random Survival Forest Machine Learning for the Prediction of Cardiovascular Events Among Patients With a Measured Lipoprotein(a) Level: A Model Development Study. *Circ Genom Precis Med*. Feb 2025;18(1):e004629.
27. Morris MC, Moradi H, Aslani M, Sims M, Schlundt D, Kouros CD, et al. Predicting incident cardiovascular disease among African-American adults: A deep learning approach to evaluate social determinants of health in the Jackson heart study. *PLOS ONE*. Nov 10, 2023;18(11):e0294050. doi:10.1371/journal.pone.0294050
28. Wang JX, Liang MM, Lu PJ, Cui Z, Liang Y, Wang YH, et al. Using Machine Learning to Predict MACEs Risk in Patients with Premature Myocardial Infarction. *Rev Cardiovasc Med*. May 2025;26(5):31298. doi:10.31083/RCM31298 PubMed PMID: 40475716; PubMed Central PMCID: PMC12135653.

29. Yan Q, Liu G, Wang R, Li D, Chen X, Cong J, et al. Explainable machine learning algorithm to predict cardiovascular event in patients undergoing peritoneal dialysis. *BMC Med Inform Decis Mak.* Apr 22, 2025;25(1):172. doi:10.1186/s12911-025-03003-w
30. Segar MW, Vaduganathan M, Patel KV, McGuire DK, Butler J, Fonarow GC, et al. Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. *Diabetes Care.* Dec 2019;42(12):2298-306. doi:10.2337/dc19-0587 PubMed PMID: 31519694; PubMed Central PMCID: PMC7364669.
31. Matheson MB, Kato Y, Baba S, Cox C, Lima JAC, Ambale-Venkatesh B. Cardiovascular Risk Prediction Using Machine Learning in a Large Japanese Cohort. *Circ Rep.* Dec 9, 2022;4(12):595-603. doi:10.1253/circrep.CR-22-0101 PubMed PMID: 36530840; PubMed Central PMCID: PMC9726526.
32. Chhoa H, Chabriat H, Chevret S, Biard L. Comparison of models for stroke-free survival prediction in patients with CADASIL. *Sci Rep.* Dec 17, 2023;13(1):22443. doi:10.1038/s41598-023-49552-w
33. Hathaway QA, Yanamala N, Budoff MJ, Sengupta PP, Zeb I. Deep neural survival networks for cardiovascular risk prediction: The Multi-Ethnic Study of Atherosclerosis (MESA). *Comput Biol Med.* Dec 2021;139:104983. doi:10.1016/j.combiomed.2021.104983 PubMed PMID: 34749095.
34. Zhuang XD, Tian T, Liao LZ, Dong YH, Zhou HJ, Zhang SZ, et al. Deep Phenotyping and Prediction of Long-term Cardiovascular Disease: Optimized by Machine Learning. *Can J Cardiol.* Jun 2022;38(6):774-82. doi:10.1016/j.cjca.2022.02.008 PubMed PMID: 35157988.
35. Farhadian M, Dehdar Karsidani S, Mozayanimonfared A, Mahjub H. Risk factors associated with major adverse cardiac and cerebrovascular events following percutaneous coronary intervention: a 10-year follow-up comparing random survival forest and Cox proportional-hazards model. *BMC Cardiovasc Disord.* Jan 18,

2021;21(1):38. doi:10.1186/s12872-020-01834-1 PubMed PMID: 33461487; PubMed Central PMCID: PMC7814642.

36. Liu X, Zhang J, Liu L, Tang X, Gao P. A population-based recalibration method for updating survival neural networks models for cardiovascular risk prediction in United Kingdom and China. *Journal of Clinical Epidemiology*. Sept 2025;185:111895. doi:10.1016/j.jclinepi.2025.111895
37. Teshale AB, Htun HL, Vered M, Owen AJ, Ryan J, Tonkin A, et al. Artificial intelligence improves risk prediction in cardiovascular disease. *GeroScience*. Nov 22, 2024. doi:10.1007/s11357-024-01438-z
38. Barbieri S, Mehta S, Wu B, Bharat C, Poppe K, Jorm L, et al. Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach. *International Journal of Epidemiology*. Jun 1, 2022;51(3):931-44. doi:10.1093/ije/dyab258
39. Chun M, Clarke R, Cairns BJ, Clifton D, Bennett D, Chen Y, et al. Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. *J Am Med Inform Assoc*. May 9, 2021;28(8):1719-27. doi:10.1093/jamia/ocab068 PubMed PMID: 33969418; PubMed Central PMCID: PMC8324240.
40. Deng Y, Liu L, Jiang H, Peng Y, Wei Y, Zhou Z, et al. Comparison of State-of-the-Art Neural Network Survival Models with the Pooled Cohort Equations for Cardiovascular Disease Risk Prediction. *BMC Med Res Methodol*. Jan 24, 2023;23(1):22. doi:10.1186/s12874-022-01829-w PubMed PMID: 36694118; PubMed Central PMCID: PMC9872364.
41. Blanchard M, Feuilloy M, Sabil K, Gervès-Pinquié C, Gagnadoux F, Girault JM. A Deep Survival Learning Approach for Cardiovascular Risk Estimation in Patients With Sleep Apnea. *IEEE Access*. Jan 1, 2022;PP:1-1.

42. Qian X, Keerman M, Zhang X, Guo H, He J, Maimaitijiang R, et al. Study on the prediction model of atherosclerotic cardiovascular disease in the rural Xinjiang population based on survival analysis. *BMC Public Health*. Jun 1, 2023;23(1):1041. doi:10.1186/s12889-023-15630-x
43. Trudel X, Gilbert-Ouimet M, Milot A, Duchaine CS, Vézina M, Laurin D, et al. Cohort Profile: The PROspective Québec (PROQ) Study on Work and Health. *Int J Epidemiol*. Jun 1, 2018;47(3):693-693i. doi:10.1093/ije/dyy026 PubMed PMID: 29534180; PubMed Central PMCID: PMC6005032.
44. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension*. Jun 2018;71(6):e13–e115.
45. Pearson GJ, Thanassoulis G, Anderson TJ, Barry AR, Couture P, Dayan N, et al. 2021 Canadian Cardiovascular Society Guidelines for the Management of Dyslipidemia for the Prevention of Cardiovascular Disease in Adults. *Canadian Journal of Cardiology*. Aug 2021;37(8):1129-50. doi:10.1016/j.cjca.2021.03.016
46. Tu K, Wang M, Young J, Green D, Ivers NM, Butt D, et al. Validity of administrative data for identifying patients who have had a stroke or transient ischemic attack using EMRALD as a reference standard. *Can J Cardiol*. Nov 2013;29(11):1388-94. doi:10.1016/j.cjca.2013.07.676 PubMed PMID: 24075778.
47. Tu K, Mitiku T, Lee DS, Guo H, Tu JV. Validation of physician billing and hospitalization data to identify patients with ischemic heart disease using data from the Electronic Medical Record Administrative data Linked Database (EMRALD). *Canadian Journal of Cardiology*. Aug 2010;26(7):e225-8. doi:10.1016/S0828-282X(10)70412-8

48. Fan J, Arruda-Olson AM, Leibson CL, Smith C, Liu G, Bailey KR, et al. Billing code algorithms to identify cases of peripheral artery disease from administrative data. *J Am Med Inform Assoc.* Dec 2013;20(e2):e349-54. doi:10.1136/amiajnl-2013-001827 PubMed PMID: 24166724; PubMed Central PMCID: PMC3861931.
49. Schultz SE, Rothwell DM, Chen Z, Tu K. Identifying cases of congestive heart failure from administrative data: a validation study using primary care patient records. *Chronic Dis Inj Can.* Jun 2013;33(3):160-6. PubMed PMID: 23735455.
50. Blais C, Jean S, Sirois C, Rochette L, Plante C, Larocque I, et al. Quebec Integrated Chronic Disease Surveillance System (QICDSS), an innovative approach. *Chronic Dis Inj Can.* Nov 2014;34(4):226-35. PubMed PMID: 25408182.
51. Romano E, Loffredo G, Maturo F. Random Survival Forest for Censored Functional Data [Internet]. arXiv; 2024 [cited Feb 17, 2025]. Available from: <http://arxiv.org/abs/2407.15340> doi:10.48550/arXiv.2407.15340
52. Ruysinck J, Van Der Hertten J, Houthoofd R, Ongenaes F, Couckuyt I, Gadeyne B, et al. Random Survival Forests for Predicting the Bed Occupancy in the Intensive Care Unit. *Computational and Mathematical Methods in Medicine.* 2016;2016:1-7. doi:10.1155/2016/7087053
53. Friedman JH. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics.* Oct 2001;29(5):1189-232. doi:10.1214/aos/1013203451
54. Chen Y, Jia Z, Mercola D, Xie X. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Comput Math Methods Med.* 2013;2013:873595. doi:10.1155/2013/873595 PubMed PMID: 24348746; PubMed Central PMCID: PMC3853154.
55. Nagpal C, Li X, Dubrawski A. Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data With Competing Risks. *IEEE Journal of Biomedical and Health Informatics.* Aug 2021;25(8):3163-75. doi:10.1109/JBHI.2021.3052441

56. Yang X, Qiu H, Wang L, Wang X. Predicting Colorectal Cancer Survival Using Time-to-Event Machine Learning: Retrospective Cohort Study. *J Med Internet Res*. Oct 26, 2023;25:e44417. doi:10.2196/44417 PubMed PMID: 37883174; PubMed Central PMCID: PMC10636616.
57. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. Apr 2024;e078378. doi:10.1136/bmj-2023-078378
58. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* [Internet]. New York, NY, USA: Association for Computing Machinery; 2019 [cited Feb 10, 2025]. p. 2623-31. (KDD '19). Available from: <https://doi.org/10.1145/3292500.3330701> doi:10.1145/3292500.3330701
59. Zhang F, Petersen M, Johnson L, Hall J, O'Bryant SE. Hyperparameter Tuning with High Performance Computing Machine Learning for Imbalanced Alzheimer's Disease Data. *Appl Sci (Basel)*. Jul 1, 2022;12(13):6670. doi:10.3390/app12136670 PubMed PMID: 36381541; PubMed Central PMCID: PMC9662287.
60. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *The Annals of Applied Statistics*. Sept 2008;2(3):841-60. doi:10.1214/08-AOAS169
61. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J Stat Soft*. 2017;77(1). doi:10.18637/jss.v077.i01
62. Yousefimehr B, Ghatee M, Seifi MA, Fazli J, Tavakoli S, Rafei Z, et al. Data Balancing Strategies: A Survey of Resampling and Augmentation Methods [Internet]. *arXiv*; 2025 [cited Jul 3, 2025]. Available from: <http://arxiv.org/abs/2505.13518> doi:10.48550/arXiv.2505.13518

63. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*. Dec 16, 2019;17(1):230. doi:10.1186/s12916-019-1466-7
64. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media; 2008. 508 p.
65. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* [Internet]. Cham: Springer International Publishing; 2019 [cited Aug 4, 2025]. (Statistics for Biology and Health). Available from: <http://link.springer.com/10.1007/978-3-030-16399-0> doi:10.1007/978-3-030-16399-0
66. Park SY, Park JE, Kim H, Park SH. Review of Statistical Methods for Evaluating the Performance of Survival or Other Time-to-Event Prediction Models (from Conventional to Deep Learning Approaches). *Korean J Radiol*. Oct 2021;22(10):1697-707. doi:10.3348/kjr.2021.0223 PubMed PMID: 34269532; PubMed Central PMCID: PMC8484151.
67. Van Geloven N, Giardiello D, Bonneville EF, Teece L, Ramspek CL, Van Smeden M, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. *BMJ*. May 24, 2022;e069249. doi:10.1136/bmj-2021-069249
68. D'Agostino RB, Nam BH. Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: *Handbook of Statistics*. Elsevier; 2003. p. 1–25. (Advances in Survival Analysis; vol. 23). Available from: <https://www.sciencedirect.com/science/article/pii/S0169716103230017>. doi:10.1016/S0169-7161(03)23001-7
69. Bauer MJ, Nano N, Adolf R, Will A, Hendrich E, Martinoff SA, et al. Prognostic Value of Machine Learning–based Time-to-Event Analysis Using Coronary CT Angiography in Patients with Suspected Coronary Artery Disease. *Radiology: Cardiothoracic Imaging*. Apr 2023;5(2):e220107. doi:10.1148/ryct.220107

70. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data?
71. Li Y, Sperrin M, Ashcroft DM, Staa TP van. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. *BMJ*. Nov 4, 2020;371:m3919. doi:10.1136/bmj.m3919 PubMed PMID: 33148619.
72. Sud M, Sivaswamy A, Chu A, Austin PC, Anderson TJ, Naimark DMJ, et al. Population-Based Recalibration of the Framingham Risk Score and Pooled Cohort Equations. *Journal of the American College of Cardiology*. Oct 2022;80(14):1330-42. doi:10.1016/j.jacc.2022.07.026
73. Ko DT, Sivaswamy A, Sud M, Kotrri G, Azizi P, Koh M, et al. Calibration and discrimination of the Framingham Risk Score and the Pooled Cohort Equations. *CMAJ*. Apr 27, 2020;192(17):E442-9. doi:10.1503/cmaj.190848

Abbreviations

CVD: cardiovascular disease

FRS: Framingham Risk Score

FRS-SP: Framingham Risk Score Simplified

ML: Machine Learning

RSF: Random Survival Forests

GBSA: Gradient Boosting Survival Analysis

Supplementary Material

Supplementary Methods – Definition of Cardiovascular Disease (CVD) Outcomes

Cardiovascular disease (CVD) outcomes were defined as the first occurrence of any of the following: coronary heart disease (CHD), stroke or transient ischemic attack (TIA), congestive heart failure (CHF), or peripheral artery disease (PAD). Identification of events was based on previously validated algorithms applied to linked medico-administrative health data in Quebec.

Stroke and TIA events were identified using a validated administrative algorithm compared against electronic medical record chart abstraction from the EMRALD database, considered the gold standard [46]. The optimal case definition required **either one hospital admission** with a relevant diagnostic code or **two physician billing claims within one year**. This algorithm achieved a **sensitivity of 68.0%, specificity of 98.9%, and positive predictive value (PPV) of 66.2%**. The diagnostic codes used were ICD-9: **362.3, 430, 431, 432, 434, 435, 436**; and ICD-10: **I60, I61, I63 (excluding I63.6), I64, H34.0, H34.1, G45 (excluding G45.4)**.

CHD events were identified using an administrative algorithm validated in Ontario, Canada, against chart-abstracted diagnoses from electronic medical records (EMRALD database), considered the gold standard [47]. CHD was defined as including myocardial infarction, angina, and both acute and chronic coronary syndromes. Relevant diagnostic codes were ICD-9: **410–414** and ICD-10: **I20–I25**. Coronary revascularization procedures were also included: angioplasty and coronary artery bypass graft (CABG), using Canadian Classification of Diagnostic, Therapeutic, and Surgical Procedures (CCP: **48.02, 48.03, 48.09, 481**) and Canadian Classification of Health Interventions (CCI: **1.IJ.50, 1.IJ.57.GQ, 1.IJ.76**). The case definition required **either one hospital discharge abstract** with a relevant CHD code or **two physician billing claims within one year**, with at least one from a specialist or a general practitioner in a hospital or emergency department setting. This algorithm achieved a sensitivity of **77.0%**, specificity of **98.0%**, and positive predictive value (PPV) of **78.8%**.

PAD events were identified using a diagnostic-procedural algorithm originally developed by validated against vascular laboratory assessments (ankle-brachial index ≤ 0.9 or >1.4) at the Mayo Clinic in Rochester, MN [48]. This model-based algorithm combined ICD-9-CM and Current Procedural Terminology, 4th revision (CPT-4) codes using a logistic regression model to assign weighted scores to 13 PAD-related diagnoses and procedures. A cut-off score of ≥ 8 yielded an **86% sensitivity, 83% specificity, and PPV of 91%** in the validation set. Procedure codes were selected to reflect diagnostic imaging and revascularization relevant to lower extremity arterial disease, excluding non-atherosclerotic causes when possible. In our study, this approach was further extended to incorporate procedures performed both before and after 2004. Diagnostic and procedure codes were mapped using the CCP for earlier years, and each code was assigned a weight ranging from 2 to 15, consistent with the original scoring method of Fan et al [48]. Given the lower specificity of CCP codes compared to more recent classifications, the lowest weights were assigned to CCP-based procedures. A PAD event was defined as a cumulative weighted score of ≥ 8 . The complete list of diagnosis and procedure codes, along with their assigned weights, is provided in **Supplementary Table S1**.

CHF events were identified using an administrative data algorithm validated in Ontario, Canada, against diagnoses abstracted from electronic medical records (EMRALD database), considered the gold standard [49]. The optimal algorithm required **either one hospital discharge abstract or one outpatient physician billing, followed by a second CHF-related record (from either source) within one year**. This definition achieved a **sensitivity of 84.8%, a specificity of 97.0%, and a PPV of 55.6%**. CHF was defined using a restricted set of diagnostic codes: ICD-9: **428**, and ICD-10: **I50.0, I50.1, I50.9**.

Supplementary Table S1. Diagnosis and Procedure Codes Used to Define Major Cardiovascular Outcomes (Coronary Heart Disease, Stroke/Transient Ischemic Attack, Congestive Heart Failure, and Peripheral Artery Disease), with Corresponding Validation Sources, Sensitivity, Specificity, and Positive Predictive Value

Validation Reference	Event Definition	Sn (%)	SP (%)	PPV (%)	ICD-9 Codes	DEF ICD-9 CODES	ICD-10 Codes	DEF ICD-10 Codes	CCI Codes	DEF CCI Codes	CCP Codes	DEF CCP Codes
CORONARY HEART DISEASE												
Tu et al., 2010 (EMRALD, Ontario, EMR)	≥ 1 hospitalization or ≥2 physician claims within 1 year	77	98	78,8	410	Acute myocardial infarction (MI)	I20	Angina pectoris	1.IJ.50	Coronary Artery Dilatation Procedures (Percutaneous Transluminal Approach)	48,02	Percutaneous Transluminal Coronary Artery Extraction and Dilation Procedures
					411	Other acute and subacute forms of ischemic heart disease (e.g., unstable angina)	I21	Acute myocardial infarction	1.IJ.57.GQ	Percutaneous Transluminal Coronary Artery Extraction Procedures	48,03	Percutaneous Coronary Intervention (PCI) with Stent Insertion
					412	Old myocardial infarction	I22	Subsequent myocardial infarction	1.IJ.76	Coronary Artery Bypass Grafting (CABG) Procedure Code	48,09	Coronary Artery Surgery without Graft (e.g., Endarterectomy, Atherectomy)
					413	Angina pectoris	I23	Certain current complications following acute myocardial infarction			481	Coronary Artery Bypass Grafting (CABG) Procedure Code
					414	Other forms of chronic ischemic heart disease (e.g., coronary atherosclerosis, chronic ischemic heart disease unspecified)	I24 I25	Other acute ischemic heart diseases Chronic ischemic heart disease				
STROKE OR TRANSIENT ISCHEMIC ATTACK												
Tu et al., 2013 (EMRALD, Ontario, EMR)	≥ 1 hospitalization or ≥ 2 physician claims within 1 year	68	98,9	66,2	362,3	Retinal vascular occlusion	I60	Nontraumatic subarachnoid hemorrhage				
					430	Subarachnoid hemorrhage	I61	Nontraumatic intracerebral hemorrhage				
					431	Intracerebral hemorrhage	I63 (excl. I63.6)	Cerebral infarction (excl. cerebral infarction due to cerebral venous thrombosis, nonpyogenic)				
					432	Other and unspecified intracranial hemorrhage	I64	Stroke, not specified as haemorrhage or infarction				

Validation Reference	Event Definition	Sn (%)	SP (%)	PPV (%)	ICD-9 Codes	DEF ICD-9 CODES	ICD-10 Codes	DEF ICD-10 Codes	CCI Codes	DEF CCI Codes	CCP Codes	DEF CCP Codes				
					434	Occlusion of cerebral arteries	H34.0	Transient retinal artery occlusion								
					435	Transient cerebral ischemia	H34.1	Central retinal artery occlusion								
					436	Acute, but ill-defined, cerebrovascular disease	G45 (excl. G45.4)	Transient cerebral ischemic attacks and related syndromes (excl. Transient global amnesia)								
CONGESTIVE HEART FAILURE																
Schultz et al., 2013 (EMRALD, Ontario, EMR)	≥ 1 hospitalization or ≥ 1 physician claim followed by another within 1 year	84,8	97	55,6	428	Heart failure	I50.0	Congestive heart failure								
							I50.1	Left ventricular failure, unspecified								
							I50.9	Heart failure, unspecified								
PERIPHERAL ARTERY DISEASE																
Fan et al., 2013 (Mayo Clinic Vascular Lab) adapted by Huisman et al., 2016	Cumulative diagnostic/procedural weight ≥ 8	86	83	91	WEIGHT			WEIGHT		WEIGHT		WEIGHT				
					440,2	Atherosclerosis of native arteries of the extremities, unspecified	10	I70.2	Atherosclerosis of the native arteries of the extremities	9	1WK93	Amputation of toe	9	9611	Amputation of toe	9
					440,21	Atherosclerosis of native arteries of the extremities with intermittent claudication	14				3KG10VA	Radiography of the Leg Arteries (covers both unilateral and bilateral)	7	5088	Peripheral vascular bypass	7
					440,22	Atherosclerosis of native arteries of the extremities with rest pain	9				3KG12VA					
					440,23	Atherosclerosis of native arteries of the	11				3KG10VX					

Validation Reference	Event Definition	Sn (%)	SP (%)	PPV (%)	ICD-9 Codes	DEF ICD-9 Codes	ICD-10 Codes	DEF ICD-10 Codes	CCI Codes	DEF CCI Codes	CCP Codes	DEF CCP Codes			
					440,24	extremities with ulceration Atherosclerosis of native arteries of the extremities with gangrene	15		3KG20	CT angiogram – abdominal aorta and lower extremity runoff	6	275	CT Scan of the Pelvic Vessels, Perineum, and Gluteal Region with or without Contrast	6	
					443,9	Peripheral vascular disease, unspecified	7	I73.9	Peripheral vascular disease, unspecified	7	3KG40	MRA lower extremity with or without contrast	276	MRI of the Pelvic Vessels, Perineum, and Gluteal Region With or Without Contrast	7
					440,9	Generalized and unspecified atherosclerosis	6	I70.9	Other and unspecified atherosclerosis	6	3KG30	Non-invasive physiologic studies of lower extremity arteries	287	Ultrasound of the Leg Arteries	2
Supplementary Table 1 footnote. Diagnostic and procedural codes were used to define major cardiovascular outcomes based on validated algorithms. For coronary heart disease (CHD), stroke or transient ischemic attack (TIA), and congestive heart failure (CHF), validated case definitions from Ontario administrative databases (EMRALD) were used, including combinations of ICD-9 and ICD-10 diagnostic codes and physician billing data. Peripheral artery disease (PAD) was defined using a weighted diagnostic algorithm developed by Fan et al. (2013) and adapted for Canadian data by Huisman et al. (2016). This adaptation involved mapping ICD-9-CM and CPT-4 codes to equivalent ICD-10-CA, CCI, and CCP codes. A PAD event was defined as a cumulative weighted score ≥8, with weights ranging from 2 to 15 based on the original logistic regression model. Lower weights were assigned to codes with lower specificity (e.g., CCP codes). Full details on definitions and code mappings are provided in the table. Abbreviations: CHD = coronary heart disease; TIA = transient ischemic attack; CHF = congestive heart failure; PAD = peripheral artery disease; ICD = International Classification of Diseases; CCI = Canadian Classification of Health Interventions; CCP = Canadian Classification of Diagnostic, Therapeutic, and Surgical Procedures; CPT = Current Procedural Terminology; EMRALD = Electronic Medical Record Administrative data Linked Database; Sn = sensitivity; SP = specificity; PPV = positive predictive value.															

Supplementary Table S2. Comparison of the Original Framingham Risk Score (FRS) and the Framingham-Simplified (SP) Models (D’Agostino et al., 2008) [13]

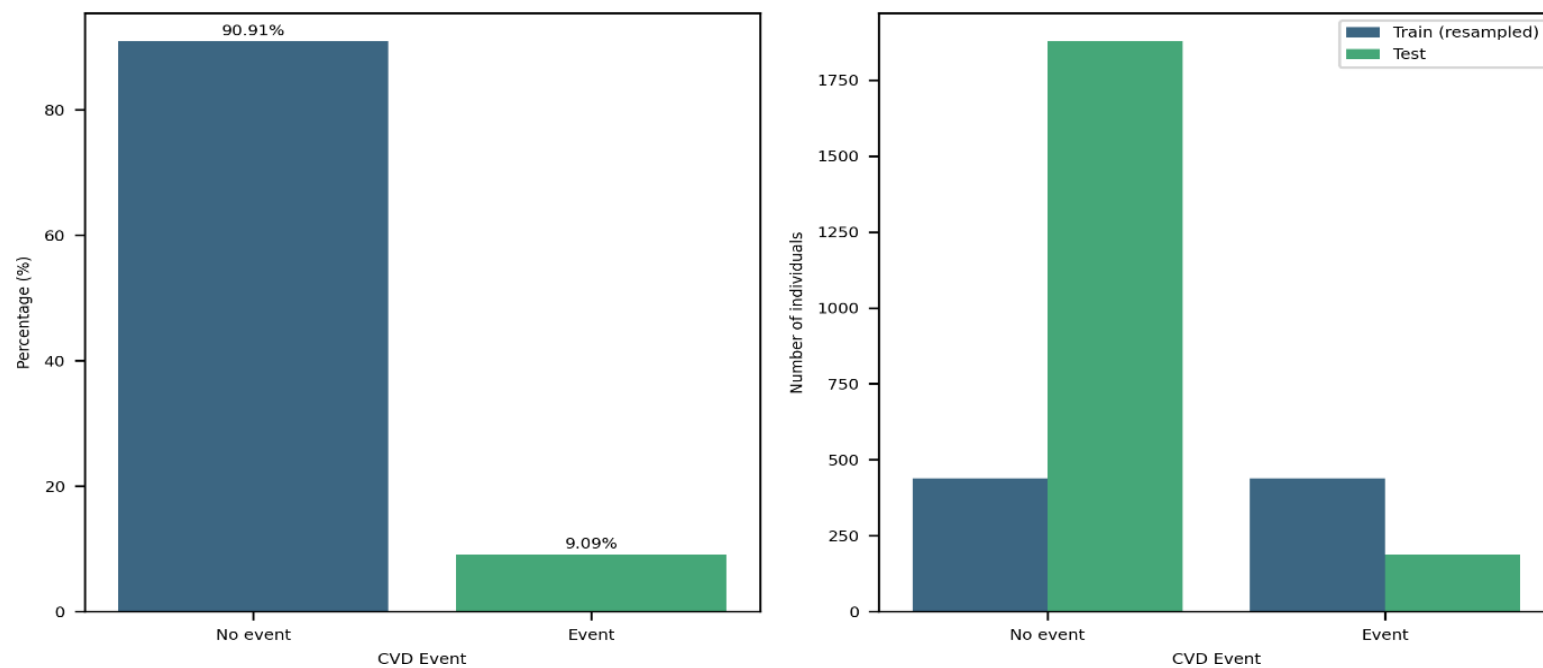
Component	Framingham Risk Score (FRS) Prediction Model	Framingham Simplified (FRS-SP) Prediction Model
Outcome predicted	First occurrence of major cardiovascular disease event, including coronary heart disease, stroke, transient ischemic attacks, heart failure, and peripheral artery disease	Same outcome predicted as FRS
Prediction horizon	10 years	10 years
Population of derivation	Framingham Offspring cohort, U.S., aged 30-74 years	Same population of derivation as FRS
Model type	Sex-specific Cox proportional hazards models	Same model type as FRS
Predictors included	- Age (years) - Total cholesterol (mg/dL) - HDL cholesterol (mg/dL) - SBP (mm Hg) - Treatment for hypertension (yes/no) - Smoking status (yes/no) - Diabetes (yes/no)	- Age (years) - BMI (kg/m²) - SBP (mm Hg) - Treatment for hypertension (yes/no) - Smoking status (yes/no) - Diabetes (yes/no)
Key difference	Includes lipid profile (total and HDL cholesterol)	BMI replaces cholesterol to simplify use in primary care settings
Risk categories (10-year risk)	- Low: <10% - Intermediate: 10-19.9% - High: ≥20%	Same risk category thresholds as FRS
Source: Adapted from D’Agostino RB Sr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. <i>Circulation</i> . 2008;117(6):743–753.		

Supplementary Table S3. Optimized hyper-parameters for OPTUNA

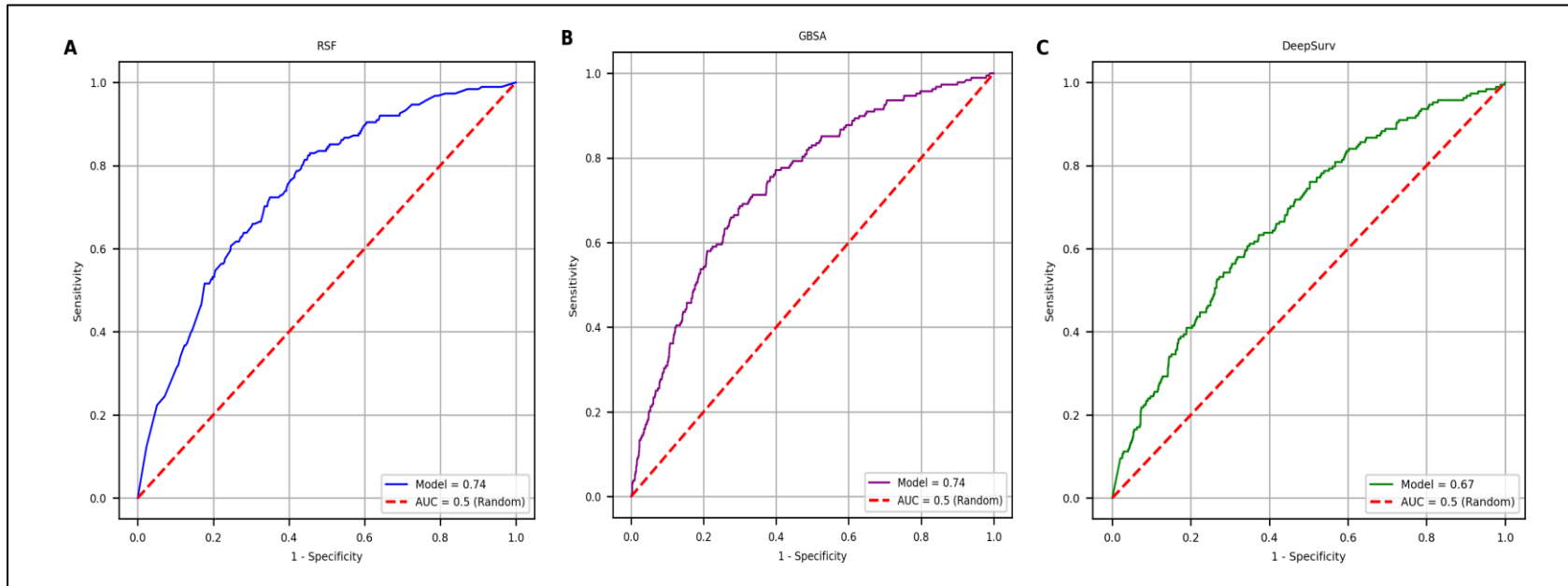
Model	Parameter	Search Range	Optimal Value
RSF	n_estimators	[100, 1000]	385
	max_depth	[5, 100]	76
	max_features	[0.3, 1.0]	0.990
	min_samples_split	[50, 200] (step=10)	190
	min_samples_leaf	[50, 200] (step=10)	90
	min_weight_fraction_leaf	[0.05, 0.2] (step=0.01)	0.09
	max_leaf_nodes	[10, 1000]	17
	bootstrap	[True, False]	False
GBSA	n_estimators	[100, 1000]	782
	learning_rate	[0.001, 0.5] (log)	0.210
	subsample	[0.5, 1.0]	0.996
	Criterion	["friedman_mse", "squared_error"]	friedman_mse
	min_samples_split	[50, 200] (step=10)	140
	min_samples_leaf	[1, 50]	36
	min_weight_fraction_leaf	[0.05, 0.2] (step=0.01)	0.13
	max_depth	[5, 100]	88
	min_impurity_decrease	[0.0, 0.05]	0.038
	max_features	[0.3, 1.0]	0.300
	max_leaf_nodes	[10, 1000]	772
	validation_fraction	[0.1, 0.5]	0.243
DeepSurv	n_iter_no_change	[10, 50]	21
	batch_size	[64, 128, 256, 512, 1024]	128
	learning_rate	[1e-5, 5e-3] (log)	0.00017
	num_layers	[1, 5]	1

	num_nodes_layer_0	[64, 128, 256]	64
	dropout	[0.1, 0.3, 0.5]	0.1
	epochs	[50, 500]	331
	optimizer	["adam", "sgd", ..., "adadelata", "nadam"]	adadelata
	activation	["relu", "tanh", "leaky_relu", "elu"]	leaky_relu
	weight_decay	[0.1, 0.05, 0.02, 0.01]	0.05
	patience	[5, 50]	46
	batch_size	[64, 128, 256, 512, 1024]	128
	learning_rate	[1e-5, 5e-3] (log)	0.00017
	num_layers	[1, 5]	1

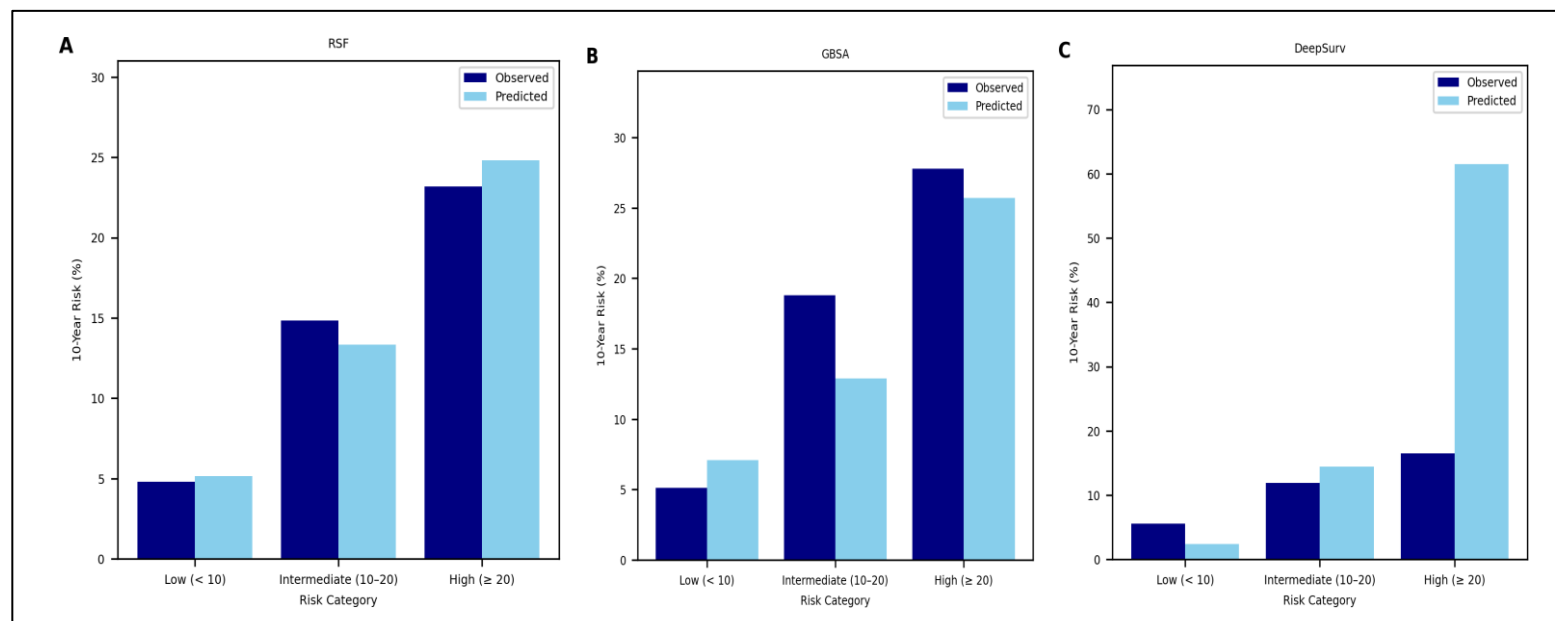
Supplementary Figure S1. Distribution of CVD events before and after under sampling



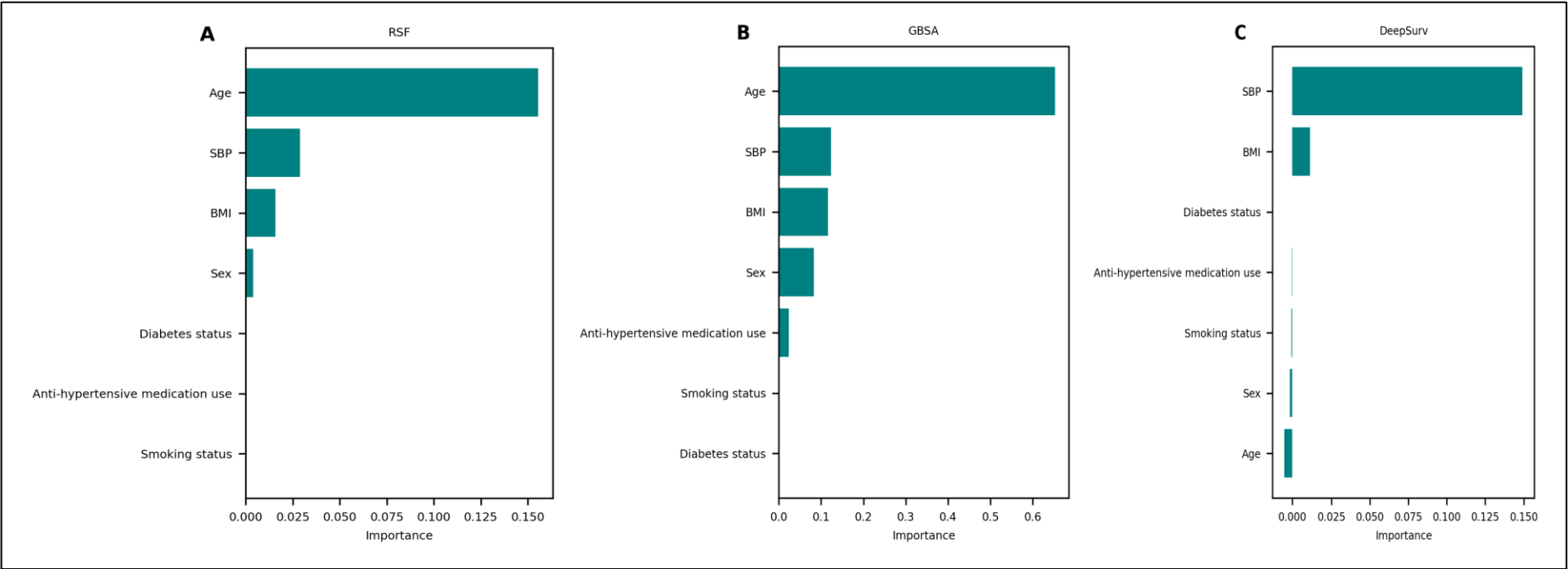
Supplementary Figure S2. Receiver Operating Characteristic Curves for Cardiovascular Disease Risk Prediction Models (Overall Population)



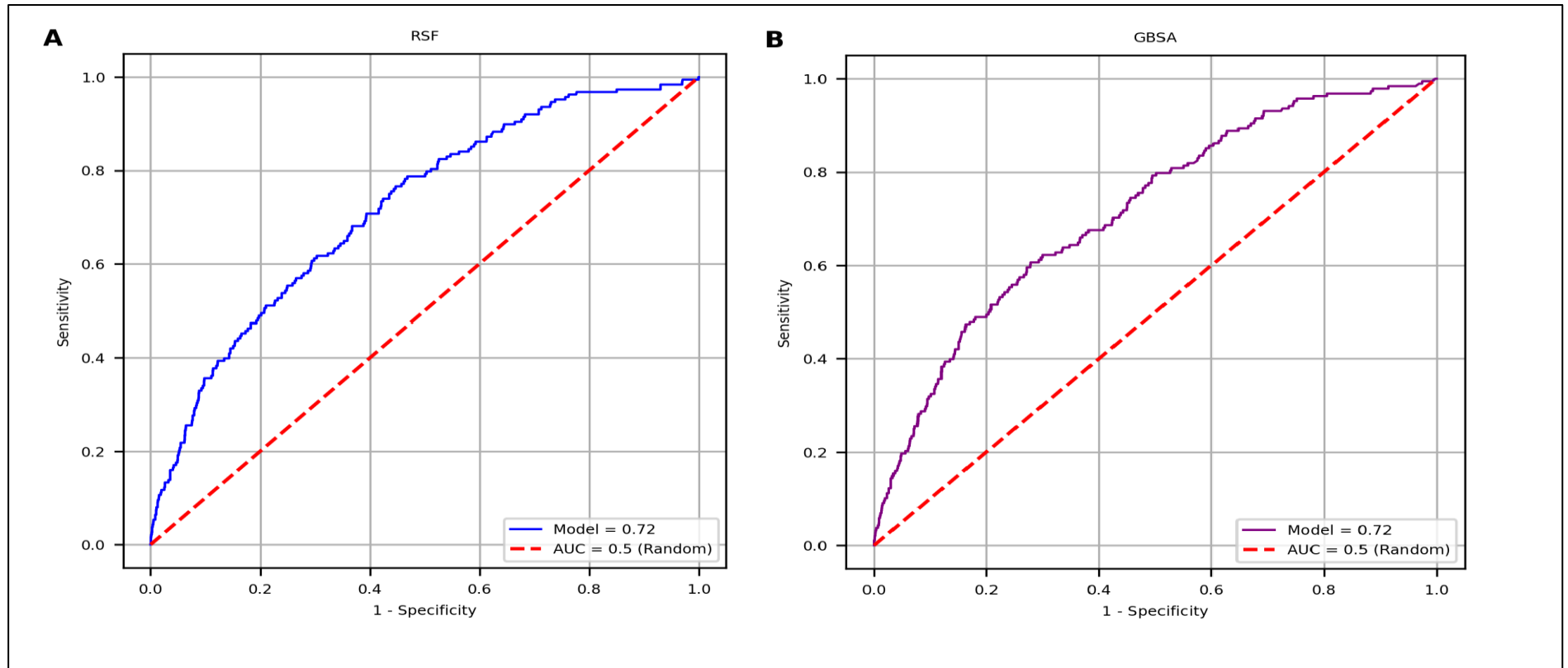
Supplementary Figure S3. Predicted and Observed 10-Year Cardiovascular Disease Risk by Risk Category (Overall population)



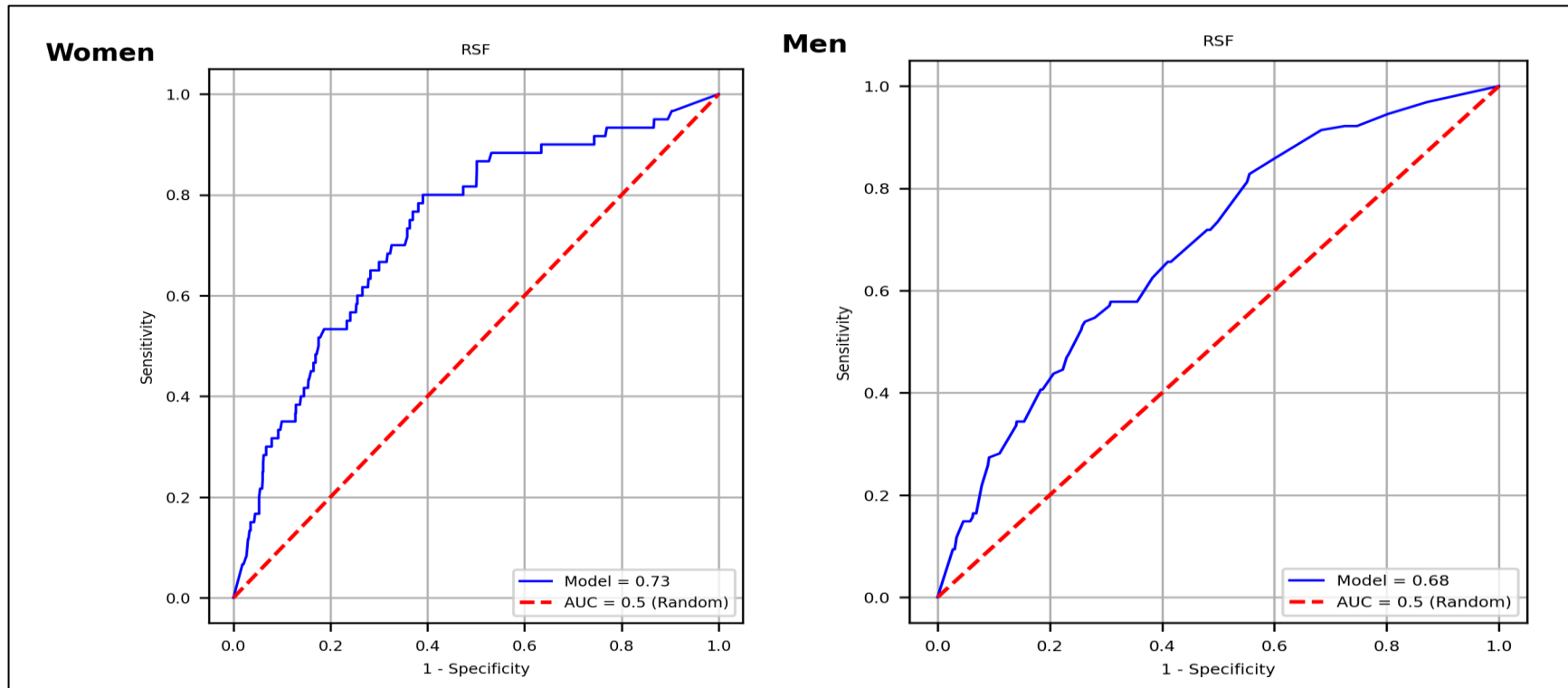
Supplementary Figure S4. Comparative Analysis of Variable Contributions to Cardiovascular Risk Prediction Models (Overall population)



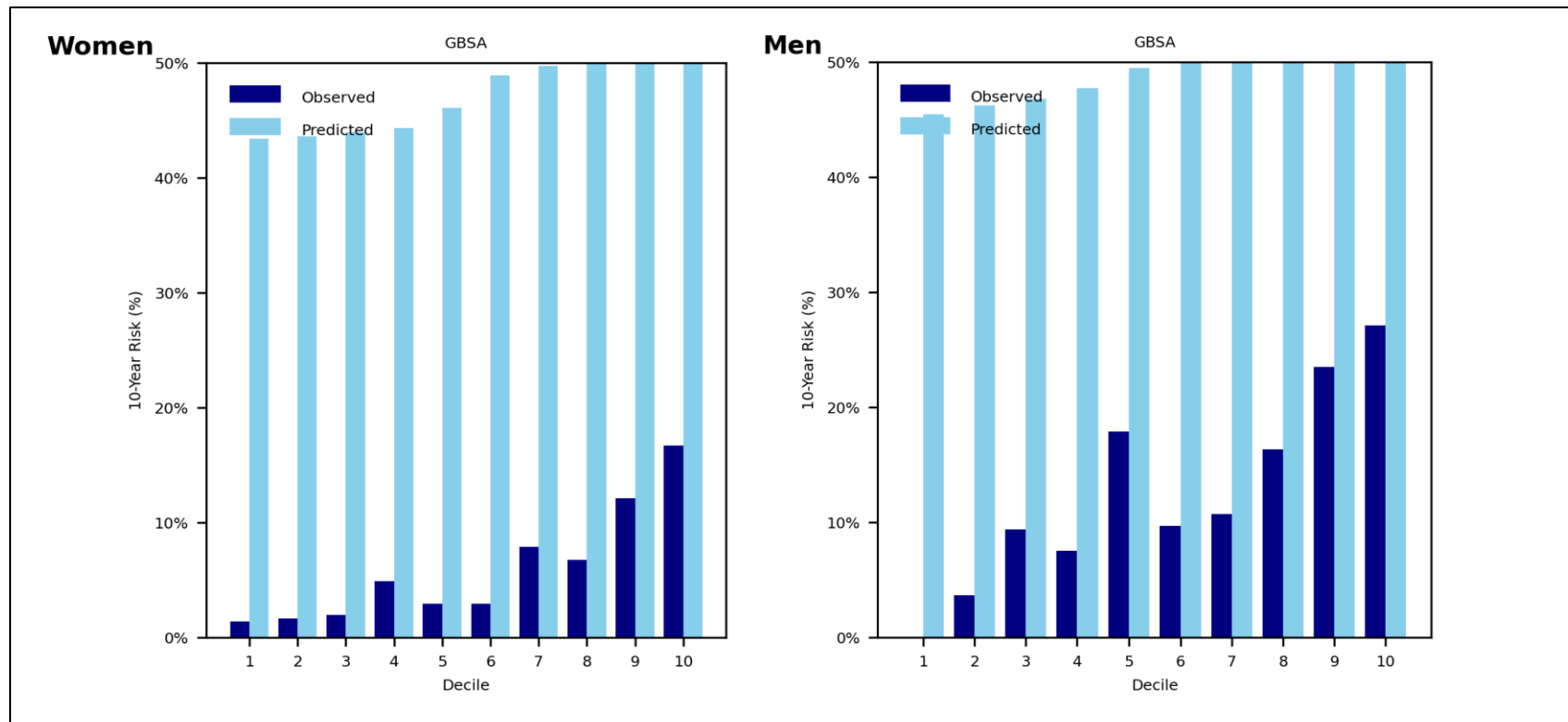
Supplementary Figure S5. Receiver Operating Characteristic Curves for Cardiovascular Disease Risk Prediction Models (Overall Population) with RandomUnderSampling



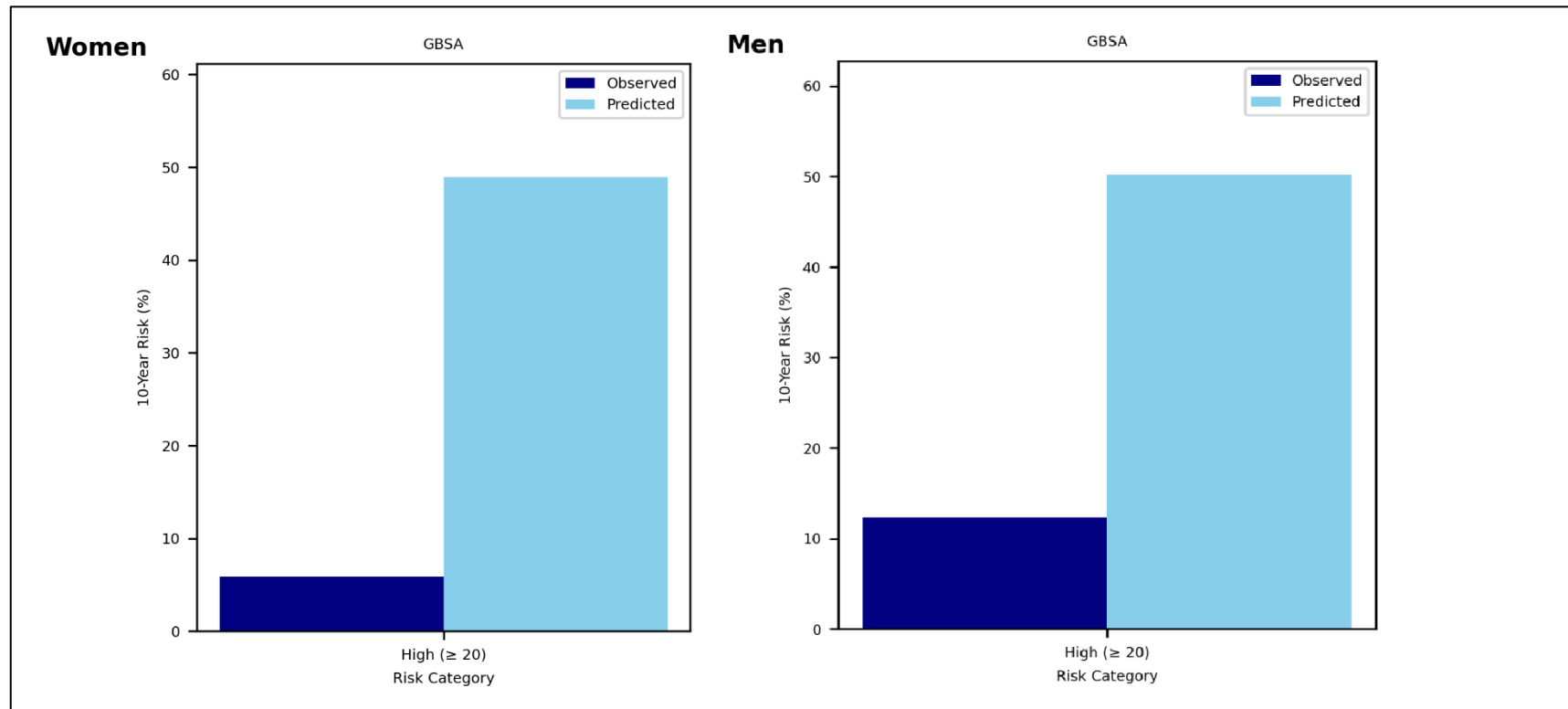
Supplementary Figure S6. Receiver Operating Characteristic Curves for Cardiovascular Disease Risk Prediction Models (Men and Women) with RandomUnderSampling for Random Survival Forests (RSF)



Supplementary Figure S7. Comparison of Predicted and Observed 10-Year CVD Risk Across Deciles of Risk (Women and Men)
RandomUnderSampling for Gradient Boosting Survival Analysis (GBSA)



Supplementary Figure S8. Predicted and Observed 10-Year Cardiovascular Disease Risk by Risk Category (Men and Women) with RandomUnderSampling for Gradient Boosting Survival Analysis (GBSA)



CONCLUSION GÉNÉRALE

2.2 SYNTHÈSE DES RESULTATS

Ce mémoire avait pour objectif d'évaluer et de comparer la performance de trois (3) modèles d'AA appliqués à l'analyse de survie (RSF, GBSA et DeepSurv) pour la prédiction du risque cardiovasculaires sur 10 ans, en s'appuyant sur des prédicteurs issus d'outils cliniques existants. Les modèles RSF et GBSA ont montré les meilleures capacités discriminatives (C-index=0.73 et AUC=0.74). là où l'approche par réseau de neurones profond (DeepSurv) a montré ses limites à calibrer correctement le risque. La calibration moyenne était de 1.02 pour RSF et 1.03 pour GBSA, représentant une légère surestimation du risque de 2 % et de 3 %. Un faible score de Brier a été obtenu (RSF = 0.0779 et GBSA = 0.0785), qui tient compte à la fois de la calibration et de la discrimination, a également indiqué une précision globale de la prédiction cohérente avec celle des modèles cliniques bien calibrés (Steyerberg, 2019). Dans notre étude, les performances observées sont globalement comparables à celles rapportées dans les travaux de Ko et al (Ko et al., 2020) et de Sud et al (Sud et al., 2022), tant en termes de discrimination que de calibration, malgré des différences méthodologiques, notamment la durée de suivi et les approches de modélisation. Au-delà des performances observées, cette recherche met en évidence l'importance des choix méthodologiques dans la gestion des données déséquilibrées. Les analyses révèlent que le réglage précis des hyperparamètres contribue davantage au maintien d'une bonne calibration des modèles que les approches de sous-échantillonnage, lesquelles peuvent, dans certains cas, altérer la qualité prédictive. Néanmoins, le déploiement de ces modèles en contexte clinique devra s'accompagner d'une vigilance accrue concernant leurs performances au sein des sous-populations féminines à haut risque, suggérant la nécessité d'améliorations futures, notamment par l'élargissement et l'enrichissement des cohortes disponibles.

L'importance d'une analyse stratifiée par sexe et par catégories de risque a été soulignée, ainsi des différences de performance persistent entre les sous-groupes, suggérant

que les modèles peuvent se comporter différemment selon le profil de risque des individus. Ces résultats appuient la nécessité d'intégrer des analyses de sous-groupes dans le développement et l'évaluation des modèles prédictifs afin d'améliorer l'équité des prédictions. Enfin, à plus long terme, ce travail ouvre la voie à une réflexion sur l'implantation progressive des modèles de survie basés sur l'apprentissage automatique comme outils d'aide à la décision clinique.

2.3 LIMITATIONS

Malgré, les bons résultats obtenus certaines limites doivent être soulignées. La cohorte étant composée majoritairement de participants en bonne santé, la généralisabilité des modèles pourrait être restreinte. En effet, si ces modèles discriminent efficacement à l'échelle populationnelle, leur précision risque d'être compromise lorsqu'ils sont appliqués à des populations présentant davantage de comorbidités.

Ainsi, il est reconnu que la nature « boîte noire » des algorithmes d'AA, notamment des réseaux de neurones, peut rendre leur interprétation difficile. Ceci tient à la complexité inhérente des interactions entre les variables de risque et à leurs effets indépendants sur le résultat. De plus, la sous-représentation des femmes appartenant aux catégories de risque élevé, limite l'évaluation robuste de la calibration pour ce sous-groupe spécifique. Cette contrainte souligne l'importance de disposer de cohortes plus diversifiées et mieux équilibrées afin de garantir une évaluation équitable des performances des modèles.

2.4 OBJECTIFS FUTURS

Les perspectives de recherche issues de ce travail sont multiples. Sur le plan méthodologique, l'utilisation d'architectures de Deep Learning plus avancées, telles que les modèles temporels (RNN, LSTM), pourrait permettre de mieux capturer les relations complexes entre les facteurs de risque. L'intégration de techniques d'explicabilité (XAI), telles que SHAP ou LIME, serait également explorée afin d'améliorer l'interprétabilité des modèles et d'en renforcer l'acceptabilité clinique.

En plus de la stratification par sexe et par catégories de risque, des analyses plus approfondies pourraient être menées en évaluant les performances des modèles au sein de différents sous-groupes de population, notamment selon l'âge, le statut socioéconomique ou d'autres caractéristiques démographiques, afin d'identifier d'éventuelles disparités de performance. Par ailleurs, au-delà de l'augmentation de la taille de l'échantillon, une ingénierie des fonctionnalités (*feature engineering*) plus avancée pourrait être mise en œuvre, intégrant des facteurs psychosociaux, socioéconomiques et environnementaux. Ces variables pourraient inclure, par exemple, la durée et l'intensité des heures de travail, les charges familiales, le niveau de stress psychosocial, ainsi que d'autres déterminants du mode de vie. L'intégration de ces facteurs pourrait permettre de mieux refléter la complexité multifactorielle du risque cardiovasculaire et d'améliorer la capacité prédictive et la pertinence clinique des modèles.

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna : A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, 2623-2631. <https://doi.org/10.1145/3292500.3330701>
2. Ambale-Venkatesh, B., Yang, X., Wu, C. O., Liu, K., Hundley, W. G., McClelland, R., Gomes, A. S., Folsom, A. R., Shea, S., Guallar, E., Bluemke, D. A., & Lima, J. A. C. (2017). Cardiovascular Event Prediction by Machine Learning. *Circulation Research*, 121(9), 1092-1101.
3. Andersson, C., Naylor, M., Tsao, C. W., Levy, D., & Vasan, R. S. (2021). Framingham Heart Study: JACC Focus Seminar, 1/8. *Journal of the American College of Cardiology*, 77(21), 2680-2692.
4. Arnould, L., Meriaudeau, F., Guenancia, C., Germanese, C., Delcourt, C., Kawasaki, R., Cheung, C. Y., Creuzot-Garcher, C., & Grzybowski, A. (2023). Using Artificial Intelligence to Analyse the Retinal Vascular Network: The Future of Cardiovascular Risk Assessment Based on Oculomics? A Narrative Review. *Ophthalmol Ther*, 12(2), 657-674. <https://doi.org/10.1007/s40123-022-00641-5>
5. Artigao-Rodenas, L. M., Carbayo-Herencia, J. A., Divisón-Garrote, J. A., Gil-Guillén, V. F., Massó-Orozco, J., Simarro-Rueda, M., Molina-Escribano, F., Sanchis, C., Carrión-Valero, L., Coca, E. L. de, Caldevilla, D., López-Abril, J., Carratalá-Munuera, C., Lopez-Pineda, A., & Albacete (GEVA), on behalf of the G. de E. V. de. (2013). Framingham Risk Score for Prediction of Cardiovascular Diseases: A Population-Based Study from Southern Europe. *PLOS ONE*, 8(9), e73529. <https://doi.org/10.1371/journal.pone.0073529>

6. Austin, P. C., & Steyerberg, E. W. (2019). The Integrated Calibration Index (ICI) and related metrics. *Statistics in Medicine*, 38(21), 4051-4065. <https://doi.org/10.1002/sim.8281>
7. Bao, L., Wang, Y., Zhuang, J., Liu, A., Dong, Y., Chu, B., Chen, X., Lu, M., Shi, L., Gao, S., Fang, L., Xiang, Q., & Ding, Y. (2022). Machine learning–based overall survival prediction of elderly patients with multiple myeloma from multicentre real-life data. *Frontiers in Oncology*, 12. <https://doi.org/10.3389/fonc.2022.922039>
8. Barbieri, S., Mehta, S., Wu, B., Bharat, C., Poppe, K., Jorm, L., & Jackson, R. (2022). Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach. *International Journal of Epidemiology*, 51(3), 931–944. <https://doi.org/10.1093/ije/dyab258>
9. Baudin, B., Cohen, A., Berthelot-Garcias, E., Meuleman, C., Dufaitre, G., Ederhy, S., Haddour, N., Boccara, F., & Cohen, A. (2009). Données épidémiologiques des maladies cardiovasculaires et prise en charge des accidents cardiovasculaires. *Revue Francophone des Laboratoires*, 2009(409), 27–39. [https://doi.org/10.1016/S1773-035X\(09\)70198-4](https://doi.org/10.1016/S1773-035X(09)70198-4)
10. Bauer, M. J., Nano, N., Adolf, R., Will, A., Hendrich, E., Martinoff, S. A., & Hadamitzky, M. (2023). Prognostic value of machine learning–based time-to-event analysis using coronary CT angiography in patients with suspected coronary artery disease. *Radiology: Cardiothoracic Imaging*, 5(2), e220107. <https://doi.org/10.1148/ryct.220107>
11. Blacher, J., & Agnoletti, D. (2011). Échelles de risque cardiovasculaire. *Médecine des Maladies Métaboliques*, 5(3), 287–293.

12. Blanchard, M., Feuilloy, M., Sabil, K., Gervès-Pinquié, C., Gagnadoux, F., & Girault, J.-M. (2022). A deep survival learning approach for cardiovascular risk estimation in patients with sleep apnea. *IEEE Access*.
13. Bradshaw, T. J., Huemann, Z., Hu, J., & Rahmim, A. (2023). A guide to cross-validation for artificial intelligence in medical imaging. *Radiology: Artificial Intelligence*, 5(4), e220232. <https://doi.org/10.1148/ryai.220232>
14. Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4).
15. World Health Organization.(2022). Cardiovascular diseases (CVDs). Retrieved November 28, 2025, from <https://www.who.int/news-room/fact->
16. Chhoa, H., Chabriat, H., Chevret, S., & Biard, L. (2023). Comparison of models for stroke-free survival prediction in patients with CADASIL. *Scientific Reports*, 13(1), 22443. <https://doi.org/10.1038/s41598-023-49552-w>
17. Cho, S. M., Austin, P. C., Ross, H. J., Abdel-Qadir, H., Chicco, D., Tomlinson, G., Taheri, C., Foroutan, F., Lawler, P. R., Billia, F., Gramolini, A., Epelman, S., Wang, B., & Lee, D. S. (2021). Machine learning compared with conventional statistical models for predicting myocardial infarction readmission and mortality: A systematic review. *Canadian Journal of Cardiology*, 37(8), 1207–1214. <https://doi.org/10.1016/j.cjca.2021.02.020>
18. Chun, M., Clarke, R., Cairns, B. J., Clifton, D., Bennett, D., Chen, Y., Guo, Y., Pei, P., Lv, J., Yu, C., Yang, L., Li, L., Chen, Z., & Zhu, T. (2021). Stroke risk prediction using machine learning: A prospective cohort study of 0.5 million Chinese adults. *Journal of the American Medical Informatics Association*, 28(8), 1719–1727. <https://doi.org/10.1093/jamia/ocab068>
19. Conroy, R. M., Pyörälä, K., Fitzgerald, A. P., Sans, S., Menotti, A., De Backer, G., De Bacquer, D., Ducimetière, P., Jousilahti, P., Keil, U., Njølstad, I.,

- Oganov, R. G., Thomsen, T., Tunstall-Pedoe, H., Tverdal, A., Wedel, H., Whincup, P., Wilhelmsen, L., Graham, I. M., & SCORE Project Group. (2003). Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *European Heart Journal*, 24(11), 987–1003.
20. Cybulska, B., & Kł osiewicz-Latoszek, L. (2019). Landmark studies in coronary heart disease epidemiology. The Framingham Heart Study after 70 years and the Seven Countries Study after 60 years. *Kardiol Pol*, 77(2), 173-180. <https://doi.org/10.5603/KP.a2019.0017>
 21. D'Agostino, R. B., & Nam, B.-H. (2003). Evaluation of the Performance of Survival Analysis Models : Discrimination and Calibration Measures. In *Handbook of Statistics* (Vol. 23, p. 1-25). Elsevier.
 22. D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care : The Framingham Heart Study. *Circulation*.
 23. Damen, J. A. A. G., Hooft, L., Schuit, E., Debray, T. P. A., Collins, G. S., Tzoulaki, I., Lassale, C. M., Siontis, G. C. M., Chiocchia, V., Roberts, C., Schlüssel, M. M., Gerry, S., Black, J. A., Heus, P., van der Schouw, Y. T., Peelen, L. M., & Moons, K. G. M. (2016). Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*, 353, i2416. <https://doi.org/10.1136/bmj.i2416>
 24. Darabi, P., Gharibzadeh, S., Khalili, D., Bagherpour-Kalo, M., & Janani, L. (2024). Optimizing cardiovascular disease mortality prediction : A super learner approach in the tehran lipid and glucose study. *BMC Medical Informatics and Decision Making*, 24(1), 97. <https://doi.org/10.1186/s12911-024-02489-0>

25. Dawber, T. R., Kannel, W. B., & Lyell, L. P. (1963). An Approach to Longitudinal Studies in a Community : The Framingham Study. *Annals of the New York Academy of Sciences*, 107(2), 539-556.
26. Dolezalova, N., Reed, A. B., Despotovic, A., Obika, B. D., Morelli, D., Aral, M., & Plans, D. (2021). Development of an accessible 10-year Digital CARDioVAscular (DiCAVA) risk assessment : A UK Biobank study. *European Heart Journal. Digital Health*, 2(3), 528-538.
27. Fan, J., Arruda-Olson, A. M., Leibson, C. L., Smith, C., Liu, G., Bailey, K. R., & Kullo, I. J. (2013). Billing code algorithms to identify cases of peripheral artery disease from administrative data. *Journal of the American Medical Informatics Association : JAMIA*, 20(e2), e349-e354.
28. Faraggi, D., & Simon, R. (1995). A neural network model for survival data. *Statistics in Medicine*, 14(1), 73-82. <https://doi.org/10.1002/sim.4780140108>
29. Farhadian, M., Dehdar Karsidani, S., Mozayanimonfared, A., & Mahjub, H. (2021). Risk factors associated with major adverse cardiac and cerebrovascular events following percutaneous coronary intervention: A 10-year follow-up comparing random survival forest and Cox proportional-hazards model. *BMC Cardiovascular Disorders*, 21(1), 38. <https://doi.org/10.1186/s12872-020-01834-1>
30. Friedman, J. H. (2001). Greedy function approximation : A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232.
31. Gautam, N., Mounsey, J. P., Yeh, E. T. H., & Al'Aref, S. J. (2023). Promises and challenges of machine learning for device therapy in heart failure. *European Heart Journal*, 44(18), 1583-1585.
32. Gilbert-Ouimet, M., Sultan-Taïeb, H., Charif, A. B., Brisson, C., Lavigne-Robichaud, M., Milot, A., Trudel, X., Demers, É., & Guertin, J. R. (2024).

Direct medical costs of cardiovascular diseases: Do cost components vary according to sex and age? *PLOS ONE*, 19(10), e0311599.

33. Grover, S. A., & Lowensteyn, I. (2011). The challenges and benefits of cardiovascular risk assessment in clinical practice. *The Canadian Journal of Cardiology*, 27(4), 481-487. <https://doi.org/10.1016/j.cjca.2011.04.008>
34. Hathaway, Q. A., Yanamala, N., Budoff, M. J., Sengupta, P. P., & Zeb, I. (2021). Deep neural survival networks for cardiovascular risk prediction : The Multi-Ethnic Study of Atherosclerosis (MESA). *Computers in Biology and Medicine*, 139, 104983.
35. Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., & Brindle, P. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom : Prospective open cohort study. *BMJ (Clinical Research Ed.)*, 335(7611), 136.
36. Hirai, H., Asahi, K., Yamaguchi, S., Mori, H., Satoh, H., Iseki, K., Moriyama, T., Yamagata, K., Tsuruya, K., Fujimoto, S., Narita, I., Konta, T., Kondo, M., Shibagaki, Y., Kasahara, M., Watanabe, T., & Shimabukuro, M. (2019). New risk prediction model of coronary heart disease in participants with and without diabetes : Assessments of the Framingham risk and Suita scores in 3-year longitudinal database in a Japanese population. *Scientific Reports*, 9(1), Article 1. <https://doi.org/10.1038/s41598-019-39049-w>
37. Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, 7(3), 355-373.
38. Ishii, M., Kaikita, K., Yasuda, S., Akao, M., Ako, J., Matoba, T., Nakamura, M., Miyauchi, K., Hagiwara, N., Kimura, K., Hirayama, A., Nishihara, E., Nakamura, S., Matsui, K., Ogawa, H., Tsujita, K., & AFIRE Investigators.

- (2023). Risk prediction score for clinical outcome in atrial fibrillation and stable coronary artery disease. *Open Heart*, 10(1), e002292.
39. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3), 841-860. <https://doi.org/10.1214/08-AOAS169>
 40. Ishwaran, H., & Lu, M. (2019). Random survival forests. In W. Wei (Ed.), *Wiley StatsRef: Statistics Reference Online*, 1–13.
 41. Jain, R., Stone, J. A., Agarwal, G., Andrade, J. G., Bacon, S. L., Bajaj, H. S., Baker, B., Cheng, G., Dannenbaum, D., Gelfer, M., Habert, J., Hickey, J., Keshavjee, K., Kitty, D., Lindsay, P., L'Abbé, M. R., Lau, D. C. W., Macle, L., McDonald, M., ... Tobe, S. W. (2023). Ligne directrice C-CHANGE pour l'harmonisation des lignes directrices nationales de prévention et de prise en charge des maladies cardiovasculaires en contexte de soins primaires au Canada : Mise à jour 2022. *CMAJ : Canadian Medical Association Journal*, 195(1), E21-E42. <https://doi.org/10.1503/cmaj.220138-f>
 42. Jegan, R., & Nimi, W. S. (2023). On the development of low power wearable devices for assessment of physiological vital parameters: A systematic review. *Zeitschrift Fur Gesundheitswissenschaften = Journal of Public Health*, 1-16.
 43. Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv : Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 24.
 44. Ko, D. T., Sivaswamy, A., Sud, M., Kottri, G., Azizi, P., Koh, M., Austin, P. C., Lee, D. S., Roifman, I., Thanassoulis, G., Tu, K., Udell, J. A., Wijeyesundera, H. C., & Anderson, T. J. (2020). Calibration and discrimination of the

Framingham Risk Score and the Pooled Cohort Equations. *Canadian Medical Association Journal*, 192(17), E442-E449.

45. Liu, X., Zhang, J., Liu, L., Tang, X., & Gao, P. (2025). A population-based recalibration method for updating survival neural networks models for cardiovascular risk prediction in United Kingdom and China. *Journal of Clinical Epidemiology*, 185, 111895.
46. Lloyd-Jones, D. M. (2010). Cardiovascular risk prediction: Basic concepts, current status, and future directions. *Circulation*, 121(15), 1768-1777.
47. Lusk, J. B., O'Brien, E. C., Hammill, B. G., Li, F., Mac Grory, B., Patel, M. R., Pagidipati, N. J., & Shah, N. P. (2025). Random Survival Forest Machine Learning for the Prediction of Cardiovascular Events Among Patients With a Measured Lipoprotein(a) Level : A Model Development Study. *Circulation. Genomic and Precision Medicine*, 18(1), e004629.
48. Mahesh, B. (2019). Machine Learning Algorithms -A Review. In *International Journal of Science and Research (IJSR)* (Vol. 9).
49. Meier, T., Gräfe, K., Senn, F., Sur, P., Stangl, G. I., Dawczynski, C., März, W., Kleber, M. E., & Lorkowski, S. (2019). Cardiovascular mortality attributable to dietary risk factors in 51 countries in the WHO European Region from 1990 to 2016 : A systematic analysis of the Global Burden of Disease Study. *European Journal of Epidemiology*, 34(1), 37-55.
50. Miao, F., Cai, Y.-P., Zhang, Y.-T., & Li, C.-Y. (2015). Is Random Survival Forest an Alternative to Cox Proportional Model on Predicting Cardiovascular Disease? In I. Lacković & D. Vasic (Éds.), *6th European Conference of the International Federation for Medical and Biological Engineering* (p. 740-743). Springer International Publishing. https://doi.org/10.1007/978-3-319-11128-5_184

51. Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S., & Geleijnse, G. (2021). Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific Reports*, 11(1), 6968. <https://doi.org/10.1038/s41598-021-86327-7>
52. Park, K. E., & Pepine, C. J. (2015). Assessing cardiovascular risk in women : Looking beyond traditional risk factors. *Trends in cardiovascular medicine*, 25(2), 152-153. <https://doi.org/10.1016/j.tcm.2014.10.024>
53. Park, S. Y., Park, J. E., Kim, H., & Park, S. H. (2021). Review of Statistical Methods for Evaluating the Performance of Survival or Other Time-to-Event Prediction Models (from Conventional to Deep Learning Approaches). *Korean Journal of Radiology*, 22(10), 1697-1707.
54. Pearson, G. J., Thanassoulis, G., Anderson, T. J., Barry, A. R., Couture, P., Dayan, N., Francis, G. A., Genest, J., Grégoire, J., Grover, S. A., Gupta, M., Hegele, R. A., Lau, D., Leiter, L. A., Leung, A. A., Lonn, E., Mancini, G. B. J., Manjoo, P., McPherson, R., ... Wray, W. (2021). 2021 Canadian Cardiovascular Society Guidelines for the Management of Dyslipidemia for the Prevention of Cardiovascular Disease in Adults. *Canadian Journal of Cardiology*, 37(8), 1129-1150. <https://doi.org/10.1016/j.cjca.2021.03.016>
55. Pietrobon, R., Machiavelli, A., Paulsen Rodrigues, L., Agrey, A., Nkeangnyi, L., Zechia, G., Lindenberg, A., Galvão, V., & Teixeira, L. (2025). *Supervised, Unsupervised, and Reinforcement Learning : A Conceptual Guide for Clinical Researchers* (SSRN Scholarly Paper N° 5744842). Social Science Research Network. <https://doi.org/10.2139/ssrn.5744842>
56. Pölsterl, S. (2015). *scikit-survival : A Library for Time-to-Event Analysis Built on Top of scikit-learn*.

57. Preiss, D., & Kristensen, S. L. (2015). The new pooled cohort equations risk calculator. *The Canadian Journal of Cardiology*, 31(5), 613-619.
58. Premasagar, P., Aldous, C., Esterhuizen, T. M., Gomes, B. J., Gaskell, J. W., & Tabb, D. L. (2022). Comparing conventional statistical models and machine learning in a small cohort of South African cardiac patients. *Informatics in Medicine Unlocked*, 34, 101103. <https://doi.org/10.1016/j.imu.2022.101103>
59. Qian, X., Keerman, M., Zhang, X., Guo, H., He, J., Maimaitijiang, R., Wang, X., Ma, J., Li, Y., Ma, R., & Guo, S. (2023). Study on the prediction model of atherosclerotic cardiovascular disease in the rural Xinjiang population based on survival analysis. *BMC Public Health*, 23(1), 1041.
60. Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of Conventional Statistical Methods with Machine Learning in Medicine : Diagnosis, Drug Development, and Treatment. *Medicina (Kaunas, Lithuania)*, 56(9), 455.
61. Ren, J., Liu, D., Li, G., Duan, J., Dong, J., & Liu, Z. (2022). Prediction and Risk Stratification of Cardiovascular Disease in Diabetic Kidney Disease Patients. *Frontiers in Cardiovascular Medicine*, 9, 923549.
62. Riley, R. D., Debray, T. P. A., Collins, G. S., Archer, L., Ensor, J., van Smeden, M., & Snell, K. I. E. (2021). Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine*, 40(19), 4230-4251.
63. Rodríguez, J. D., Pérez, A., & Lozano, J. A. (2010). Sensitivity analysis of kappa-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3), 569-575.
64. Romano, E., Loffredo, G., & Maturo, F. (2024, juillet). *Random Survival Forest for Censored Functional Data*. arXiv.

65. Ruyssinck, J., Van Der Hertten, J., Houthoof, R., Ongenae, F., Couckuyt, I., Gadeyne, B., Colpaert, K., Decruyenaere, J., De Turck, F., & Dhaene, T. (2016). Random Survival Forests for Predicting the Bed Occupancy in the Intensive Care Unit. *Computational and Mathematical Methods in Medicine*, 2016, 1-7. <https://doi.org/10.1155/2016/7087053>
66. Sabovčik, F., Ntalianis, E., Cauwenberghs, N., & Kuznetsova, T. (2022). Improving predictive performance in incident heart failure using machine learning and multi-center data. *Frontiers in Cardiovascular Medicine*, 9. <https://doi.org/10.3389/fcvm.2022.1011071>
67. Sah, S. (2020). *Machine Learning: A Review of Learning Types*. MATHEMATICS & COMPUTER SCIENCE.
68. Santos, J. V., Vandenberghe, D., Lobo, M., & Freitas, A. (2020). Cost of cardiovascular disease prevention: Towards economic evaluations in prevention programs. *Annals of Translational Medicine*, 8(7), 512. <https://doi.org/10.21037/atm.2020.01.20>
69. Sarker, I. H. (2021). Machine Learning : Algorithms, Real-World Applications and Research Directions. *Sn Computer Science*, 2(3), 160.
70. Schultz, S., Rothwell, D., Chen, Z., & Tu, K. (2013). Repérage des cas d'insuffisance cardiaque congestive à partir de données administratives : Étude de validation utilisant des dossiers de patients en soins primaires. *Maladies chroniques et blessures au Canada*, 33(3), 179-186.
71. Segar, M. W., Vaduganathan, M., Patel, K. V., McGuire, D. K., Butler, J., Fonarow, G. C., Basit, M., Kannan, V., Grodin, J. L., Everett, B., Willett, D., Berry, J., & Pandey, A. (2019). Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. *Diabetes Care*, 42(12), 2298-2306.

72. Shekhar, S., Bansode, A., & Salim, A. (2022). *A Comparative study of Hyper-Parameter Optimization Tools* (arXiv:2201.06433). arXiv.
73. Singh, V., Pencina, M., Einstein, A. J., Liang, J. X., Berman, D. S., & Slomka, P. (2021). Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific Reports*, *11*, 14490.
74. Smith, E. R. (2009). The Canadian Heart Health Strategy and Action Plan. *The Canadian Journal of Cardiology*, *25*(8), 451-452.
75. Smolderen, K. G., Bell, A., Lei, Y., Cohen, E. A., Steg, P. G., Bhatt, D. L., & Mahoney, E. M. (2010). One-year costs associated with cardiovascular disease in Canada : Insights from the REduction of Atherothrombosis for Continued Health (REACH) registry. *The Canadian Journal of Cardiology*, *26*(8), e297-e305. [https://doi.org/10.1016/s0828-282x\(10\)70437-2](https://doi.org/10.1016/s0828-282x(10)70437-2)
76. Song, J., Li, J., Zhao, R., & Chu, X. (2023). Developing predictive models for surgical outcomes in patients with degenerative cervical myelopathy: A comparison of statistical and machine learning approaches. *The Spine Journal*, *0*(0).
77. Steyerberg, E. W. (2008). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media.
78. Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer International Publishing.
79. Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models : A framework for traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, *21*(1), 128-138.

80. Sud, M., Sivaswamy, A., Chu, A., Austin, P. C., Anderson, T. J., Naimark, D. M. J., Farkouh, M. E., Lee, D. S., Roifman, I., Thanassoulis, G., Tu, K., Udell, J. A., Wijeyesundera, H. C., & Ko, D. T. (2022). Population-Based Recalibration of the Framingham Risk Score and Pooled Cohort Equations. *Journal of the American College of Cardiology*, 80(14), 1330-1342.
81. Sung, J. M., Cho, I.-J., Sung, D., Kim, S., Kim, H. C., Chae, M.-H., Kavousi, M., Rueda-Ochoa, O. L., Ikram, M. A., Franco, O. H., & Chang, H.-J. (2019). Development and verification of prediction models for preventing cardiovascular diseases. *PloS One*, 14(9), e0222809.
82. Teshale, A. B., Htun, H. L., Vered, M., Owen, A. J., & Freak-Poli, R. (2024). A Systematic Review of Artificial Intelligence Models for Time-to-Event Outcome Applied in Cardiovascular Disease Risk Prediction. *Journal of Medical Systems*, 48(1), 68.
83. Teshale, A. B., Htun, H. L., Vered, M., Owen, A. J., Ryan, J., Tonkin, A., & Freak-Poli, R. (2024). Artificial intelligence improves risk prediction in cardiovascular disease. *GeroScience*. <https://doi.org/10.1007/s11357-024-01438-z>
84. Teshale, A. B., Htun, H. L., Vered, M., Owen, A. J., Ryan, J., Tonkin, A., & Freak-Poli, R. (2025). Artificial intelligence improves risk prediction in cardiovascular disease. *GeroScience*, 47(4), 5497-5502.
85. Thiriet, M. (2019). Cardiovascular Risk Factors and Markers. *Vasculopathies*, 8, 91-198. https://doi.org/10.1007/978-3-319-89315-0_2
86. Trudel, X., Gilbert-Ouimet, M., Milot, A., Duchaine, C. S., Vézina, M., Laurin, D., Sultan-Taïeb, H., Brisson, C., see Appendix 1 for the complete list of investigators and collaborators, Investigators, & Collaborators. (2018). Cohort

Profile: The PROspective Québec (PROQ) Study on Work and Health. *International Journal of Epidemiology*, 47(3), 693-693i.

87. Tu, K., Mitiku, T., Lee, D. S., Guo, H., & Tu, J. V. (2010). Validation of physician billing and hospitalization data to identify patients with ischemic heart disease using data from the Electronic Medical Record Administrative data Linked Database (EMRALD). *The Canadian Journal of Cardiology*, 26(7), e225-e228.
88. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., Bossuyt, P., Collins, G. S., Macaskill, P., McLernon, D. J., Moons, K. G. M., Steyerberg, E. W., Van Calster, B., van Smeden, M., Vickers, A. J., & On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. (2019). Calibration : The Achilles heel of predictive analytics. *BMC Medicine*, 17(1), 230. <https://doi.org/10.1186/s12916-019-1466-7>
89. Van Geloven, N., Giardiello, D., Bonneville, E. F., Teece, L., Ramspek, C. L., Van Smeden, M., Snell, K. I. E., Van Calster, B., Pohar-Perme, M., Riley, R. D., Putter, H., & Steyerberg, E. (2022). Validation of prediction models in the presence of competing risks : A guide through modern methods. *BMJ*, e069249. <https://doi.org/10.1136/bmj-2021-069249>
90. van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(2), 373-440. <https://doi.org/10.1007/s10994-019-05855-6>
91. Wang, J.-X., Liang, M.-M., Lu, P.-J., Cui, Z., Liang, Y., Wang, Y.-H., Jing, A.-R., Wang, J., Zhang, M.-L., Liu, Y., Li, C.-P., & Gao, J. (2025). Using Machine Learning to Predict MACEs Risk in Patients with Premature Myocardial Infarction. *Reviews in Cardiovascular Medicine*, 26(5), 31298.

92. Wang, P., Li, Y., & Reddy, C. K. (2017). *Machine Learning for Survival Analysis : A Survey* (arXiv:1708.04649). arXiv.
93. Wenger, N. (2017). Tailoring cardiovascular risk assessment and prevention for women: One size does not fit all. *Global Cardiology Science & Practice*, 2017(1), e201701. <https://doi.org/10.21542/gcsp.2017.1>
94. Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Dennison Himmelfarb, C., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., MacLaughlin, E. J., Muntner, P., Ovbiagele, B., Smith, S. C., Spencer, C. C., Stafford, R. S., Taler, S. J., (2018). Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults : A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension*, 71(6), e13-e115.
95. Wiegerebe, S., Kopper, P., Sonabend, R., Bischl, B., & Bender, A. (2024). Deep learning for survival analysis: A review. *Artificial Intelligence Review*, 57(3), 65.
96. Woodward, M., Brindle, P., Tunstall-Pedoe, H., & SIGN group on risk estimation. (2007). Adding social deprivation and family history to cardiovascular risk assessment : The ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart (British Cardiac Society)*, 93(2), 172-176. <https://doi.org/10.1136/hrt.2006.108167>
97. Wright, M. N., & Ziegler, A. (2017). ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1). <https://doi.org/10.18637/jss.v077.i01>
98. Y C a, P., Pulabaigari, V., & B, E. (2018). Semi-supervised learning : A brief review. *International Journal of Engineering & Technology*, 7, 81.

99. Yan, Q., Liu, G., Wang, R., Li, D., Chen, X., Cong, J., & Wang, D. (2025). Explainable machine learning algorithm to predict cardiovascular events in patients undergoing peritoneal dialysis. *BMC Medical Informatics and Decision Making*, 25(1), 172.
100. Yousefimehr, B., Ghatee, M., Seifi, M. A., Fazli, J., Tavakoli, S., Rafei, Z., Ghaffari, S., Nikahd, A., Gandomani, M. R., Orouji, A., Kashani, R. M., Heshmati, S., & Mousavi, N. S. (2025). *Data Balancing Strategies: A Survey of Resampling and Augmentation Methods* (arXiv:2505.13518). arXiv. <https://doi.org/10.48550/arXiv.2505.13518>
101. Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., & Lisheng, L. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. *The Lancet*, 364(9438), 937-952.
102. Zhang, F., Petersen, M., Johnson, L., Hall, J., & O'Bryant, S. E. (2022). Hyperparameter Tuning with High Performance Computing Machine Learning for Imbalanced Alzheimer's Disease Data. *Applied sciences (Basel, Switzerland)*, 12(13), 6670.
103. Zhang, Y., Liu, B., Zhao, R., Zhang, S., Yu, X.-Y., & Li, Y. (2020). The Influence of Sex on Cardiac Physiology and Cardiovascular Diseases. *Journal of cardiovascular translational research*, 13(1), 3-13.
104. Zhao, Z., Cheng, X., Gao, Y., Tan, F., Xue, Q., Gao, S., & He, J. (2025). Predicting survival in small cell lung cancer patients undergoing various treatments: A machine learning approach. *Translational Lung Cancer Research*, 14(3), 736-748.

105. Zhuang, X., Tian, T., Liao, L., Dong, Y., Zhou, H., Zhang, S., Chen, W., Du, Z., Wang, X., & Liao, X. (2022). Deep Phenotyping and Prediction of Long-term Cardiovascular Disease: Optimized by Machine Learning. *Canadian Journal of Cardiology*, 38(6), 774-782.
106. Zores, F., & Maeker, E. (2022). Prise en charge de l'hypertension artérielle en prévention cardiovasculaire primaire chez les personnes âgées et très âgées. *NPG Neurologie - Psychiatrie - Gériatrie*, 22(127), 5-24.

