Introduction

Les maladies cardiovasculaires constituent la première cause de mortalité dans le monde, avec environ 19,8 millions de décès en 2022, soit près de 32 % de l'ensemble des décès, dont 85 % liés à l'infarctus du myocarde et à l'accident vasculaire cérébral (1). La maladie coronarienne, en particulier, représente l'une des principales composantes de ce fardeau et demeure une cause majeure de perte d'années de vie en bonne santé, avec plus de 180 millions de DALYs attribués à l'ischémie myocardique en 2019 (2). Une proportion importante de ces événements pourrait toutefois être prévenue par l'identification précoce et la prise en charge ciblée des sujets à haut risque. C'est dans cette perspective qu'a été initiée, en 1948, la *Framingham Heart Study*, large cohorte de population générale qui a profondément transformé la compréhension des déterminants du risque cardiovasculaire, en mettant en évidence le rôle de facteurs tels que l'hypertension artérielle, les dyslipidémies, le tabagisme, le diabète ou l'obésité (3).

À partir de cette cohorte, plusieurs fonctions de risque multivariées ont été proposées pour estimer le risque à 10 ans d'événements coronariens ou cardiovasculaires, reposant classiquement sur des modèles de régression logistique ou de Cox intégrant l'âge, le sexe, la pression artérielle, les lipides, le tabagisme et le diabète (4). Ces scores sont largement diffusés en pratique clinique, mais plusieurs études ont souligné des limites de transportabilité et de calibration lorsqu'ils sont appliqués à d'autres contextes que celui de Framingham, conduisant souvent à une surestimation du risque et à la nécessité de recalibrations locales (5).

Parallèlement, le développement de méthodes de modélisation plus flexibles a suscité un intérêt croissant pour des approches capables de capturer des relations non linéaires et des interactions entre facteurs de risque, dans l'objectif d'améliorer la discrimination et la calibration des modèles prédictifs cardiovasculaires (6). L'application de ces modèles à des bases cliniques réelles pose toutefois plusieurs défis méthodologiques. Un premier enjeu concerne le déséquilibre de classes : dans de nombreux jeux de données en cardiologie, les événements d'intérêt représentent une proportion minoritaire des observations (souvent < 30 %), comme c'est le cas dans l'extrait de la cohorte de Framingham utilisé ici (environ 15 % d'événements coronariens à dix ans). Dans ce contexte, un modèle qui optimiserait uniquement l'exactitude globale (*accuracy*) tendrait à privilégier la classe majoritaire (patients sans événement) et à présenter une sensibilité très faible pour la classe événement, ce qui est problématique pour la décision clinique. Des stratégies de rééquilibrage, telles que le sur-échantillonnage synthétique

(SMOTE), le sous-échantillonnage ou l'introduction de coûts différenciés, ont été proposées pour mieux prendre en compte la classe minoritaire, avec des résultats contrastés sur la discrimination, la calibration et la généralisable des modèles (7).

Un second enjeu majeur concerne la gestion des valeurs manquantes, fréquentes dans les bases de données observationnelles. L'exclusion des enregistrements incomplets réduit la puissance et peut introduire un biais lorsque les données ne sont pas manquantes complètement au hasard. Les approches contemporaines recommandent le recours à des méthodes d'imputation multivariée, en particulier l'imputation multiple par équations en chaîne (*Multiple Imputation by Chained Equations*, MICE), qui permettent de mieux préserver les distributions des covariables et les relations entre prédicteurs et évènement que les imputations univariées simples(8). L'interaction entre ces choix de prétraitement (imputation) et les stratégies de gestion du déséquilibre de classes demeure cependant peu documentée dans le domaine de la prédiction cardiovasculaire.

Enfin, au-delà des performances ponctuelles, la quantification de l'incertitude des prédictions représente un enjeu central pour l'utilisation de modèles en pratique clinique. Sur le plan global, des méthodes de ré-échantillonnage telles que le bootstrap permettent d'estimer l'incertitude autour de métriques de performance (par exemple, intervalles de confiance pour le rappel, la précision ou l'AUC), comme recommandé dans les travaux récents sur le développement et la validation des modèles de risque (9).

Sur le plan individuel, des cadres comme la prédiction conforme (*conformal prediction*) offrent la possibilité de produire, à partir de scores de probabilité, des ensembles prédictifs assortis de garanties formelles de couverture, ce qui fournit un moyen explicite de représenter l'incertitude pour chaque patient (10).

La conception de l'étude et le reporting des résultats s'appuient sur les recommandations récentes pour les modèles de prédiction clinique, en particulier les principes de transparence et de reproductibilité portés par TRIPOD et son extension TRIPOD+AI, sans détailler ici l'ensemble de la checklist (9). Les résultats seront systématiquement mis en regard des travaux publiés sur la base Framingham ou sur des jeux de données cardiovasculaires apparentés, afin de situer les performances observées et de discuter la pertinence des choix méthodologiques en épidémiologie clinique. Au-delà de la comparaison méthodologique, l'objectif ultime est de déterminer dans quelle mesure un modèle de risque bien calibré, accompagné d'une estimation explicite de son incertitude, pourrait servir d'outil d'aide à la décision pour repérer plus

finement les patients à haut risque, guider l'intensification des traitements préventifs (par exemple statines et antihypertenseurs) et structurer des stratégies de prévention primaire plus ciblées au niveau individuel et populationnel.

Objectifs de l'étude

L'objectif général de ce travail est de développer et d'évaluer des modèles de prédiction du risque de maladie cardiaque à dix ans à partir de la base Framingham, en s'appuyant sur des algorithmes d'apprentissage automatique et sur des méthodes explicites de quantification de l'incertitude, dans une perspective d'aide à la décision médicale.

De manière plus spécifique, le projet vise à :

1. **Décrire et analyser la base de données** issue de la *Framingham Heart Study*, en caractérisant la distribution des variables cliniques et biologiques, la présence de valeurs manquantes et de valeurs extrêmes, ainsi que le déséquilibre entre classes (événement vs non-événement).
2. **Appliquer et comparer plusieurs algorithmes de prédiction supervisée** (régression logistique, forêts aléatoires, gradient boosting, XGBoost, SVM) pour l'estimation du risque de maladie cardiaque à dix ans, en évaluant leurs performances selon des métriques adaptées au déséquilibre de classes (rappel, précision, F1-score, AUC-ROC, AUC-PR, Brier score).
3. **Quantifier l'incertitude associée aux performances et aux prédictions des modèles**, en combinant des approches de rééchantillonnage pour obtenir des intervalles de confiance des métriques globales, et des méthodes d'incertitude au niveau individuel (prédiction conforme) afin d'associer un niveau de confiance explicite à chaque prédiction.
4. **Discuter la pertinence des modèles obtenus dans un contexte d'aide à la décision clinique**, en mettant en balance performance, stabilité, interprétabilité et gestion explicite de l'incertitude, et en situant les résultats par rapport aux travaux publiés sur la base Framingham et sur la prédiction du risque cardiovasculaire.

Méthodes

Type d'étude et source des données

Il s'agit d'une étude de modélisation prédictive, rétrospective, basée sur un jeu de données dérivé de la Framingham Heart Study, grande cohorte de population générale initiée en 1948 et suivie de manière longitudinale. Le jeu de données utilisé comprend 4 239 individus, initialement exempts de maladie coronarienne, pour lesquels sont disponibles des informations cliniques, biologiques et comportementales, ainsi que la survenue ou non d'un événement coronarien sur une période de dix ans.

Source des données et population d'étude

Les données proviennent d'un sous-ensemble de la Framingham Heart Study, grande cohorte prospective de population générale initiée à la fin des années 1940, qui a largement contribué à l'identification des principaux facteurs de risque cardiovasculaire et au développement de fonctions de risque multivariées. Ce jeu de données, fréquemment mobilisé à des fins pédagogiques et méthodologiques, regroupe 4 239 individus et une variable de sortie binaire indiquant la survenue ou non d'un événement coronarien à dix ans. Les participants sont des adultes initialement indemnes de maladie coronarienne, pour lesquels un ensemble standard de facteurs de risque (âge, sexe, pression artérielle, profil lipidique, tabagisme, diabète, indice de masse corporelle, antécédents cardiovasculaires) a été recueilli à l'inclusion, avant un suivi de dix ans documentant la survenue éventuelle d'un événement coronarien (11,12).

Variable cible et prédicteurs

La variable cible est un indicateur binaire de survenue d'un événement coronarien à dix ans (0 = aucun événement, 1 = événement). La prévalence de la classe événement est d'environ 15 %, traduisant un déséquilibre marqué entre patients avec et sans événement (11).

Les prédicteurs retenus sont des facteurs de risque classiques issus de Framingham :

- Données démographiques : âge (années), sexe (masculin/féminin) ;
- Comportements : statut tabagique (fumeur actuel ou non), nombre de cigarettes fumées par jour ;
- Variables cliniques : pression artérielle systolique et diastolique, fréquence cardiaque, antécédent d'accident vasculaire cérébral, hypertension connue, traitement antihypertenseur ;

- Variables biologiques : cholestérol total, glycémie ;
- Facteurs métaboliques et socio-économiques : diabète, indice de masse corporelle (IMC), niveau d'éducation.

Ces variables ont été sélectionnées car elles correspondent aux facteurs de risque classiquement utilisés dans les équations de risque de Framingham et sont disponibles pour l'ensemble ou la quasi-totalité des sujets (13).

Tableau 1 – Description des variables utilisées pour la modélisation (base Framingham)

Nom (français)	**Nom original (Framingham)**	**Type**	**Unité / codage**	**Description synthétique**
Risque_chd_10ans	TenYearCHD	Binaire	0 = non, 1 = oui	Survenue d'un événement coronarien (angor, infarctus, décès coronarien) dans les 10 ans.
Age	age	Continue	Années	Âge du participant à l'inclusion.
Sexe masculin	male	Binaire	0 = femme, 1 = homme	Sexe biologique.
Fumeur actuel	currentSmoker	Binaire	0 = non, 1 = oui	Statut de fumeur actuel au moment de l'inclusion.
Cigarettes_par_jour	cigsPerDay	Continue	Cigarettes / jour	Nombre moyen de cigarettes fumées par jour chez les fumeurs.
Cholesterol_total	totChol	Continue	mg/dL	Concentration de cholestérol total.
Glucose	glucose	Continue	mg/dL	Glycémie à jeun.
Diabète	diabetes	Binaire	0 = non, 1 = oui	Diabète connu au moment de l'inclusion.
Tension systolique	sysBP	Continue	mmHg	Pression artérielle systolique.
Tension diastolique	diaBP	Continue	mmHg	Pression artérielle diastolique.
Traitement antihypertenseur	BPMeds	Binaire	0 = non, 1 = oui	Traitement antihypertenseur en cours.
Hypertension connue	prevalentHyp	Binaire	0 = non, 1 = oui	Hypertension artérielle connue (diagnostic antérieur).
Antecedent_AVC	prevalentStroke	Binaire	0 = non, 1 = oui	Antécédent d'accident vasculaire cérébral.
IMC	BMI	Continue	kg/m²	Indice de masse corporelle.
Frequence_cardiaque	heartRate	Continue	battements / minute (bpm)	Fréquence cardiaque au repos.
Niveau éducation	education	Ordinale	1–4 (échelle catégorielle)	Niveau d'éducation (si disponible dans la version du jeu de données utilisée).

Critères d'inclusion et d'exclusion

Les analyses ont été réalisées à partir d'un sous-ensemble de la Framingham Heart Study incluant uniquement des adultes âgés de 30 ans ou plus, indemnes de maladie coronarienne au moment de l'inclusion, et pour lesquels un suivi de dix ans permettait de documenter la survenue ou non d'un événement coronarien (variable risque_chd_10ans). Les participants devaient disposer au minimum des principales mesures cliniques et biologiques (indices anthropométriques, profil lipidique, pression artérielle, données de tabagisme). Les valeurs manifestement incohérentes identifiées lors de l'analyse descriptive ont été vérifiées au cas par cas, mais aucune exclusion supplémentaire n'a été réalisée au-delà des critères du jeu d'origine, les observations extrêmes mais plausibles étant conservées afin de préserver la variabilité des prédicteurs (14).

Analyse exploratoire, prétraitement et modélisation

Analyse exploratoire et qualité des données

Avant toute étape de prétraitement (imputation, standardisation, rééquilibrage), une analyse exploratoire systématique a été réalisée afin d'évaluer la qualité des données et de documenter leur structure statistique.

Distributions univariées

Pour chaque prédicteur continu (âge, IMC, cholestérol total, pression artérielle, glycémie, fréquence cardiaque), des histogrammes et densités ont été examinés, avec calcul de statistiques descriptives (moyenne, médiane, quartiles, asymétrie, aplatissement). Les distributions étaient globalement unimodales, avec quelques asymétries modérées, ne justifiant pas de transformation systématique. Un tableau récapitulatif présente les principales statistiques pour l'ensemble des variables continues.

Corrélations et multi colinéarité

Une matrice de corrélation de Pearson entre variables continues, complétée par une représentation hiérarchique, a permis d'identifier des corrélations modérées entre certains prédicteurs (notamment les pressions artérielles), sans corrélation absolue supérieure à 0,8. Le calcul des facteurs d'inflation de variance (VIF) dans un modèle logistique complet a montré des valeurs inférieures à 4 pour l'ensemble des variables, ce qui ne suggère pas de multi colinéarité problématique pour la modélisation.

Valeurs manquantes

La structure de la manquance a été décrite à l'aide de diagrammes de barres et de matrices de "missingness", mettant en évidence des proportions de données manquantes non négligeables pour certains prédicteurs (cholestérol, glycémie, variables de pression artérielle et de tabagisme), selon des patrons hétérogènes. L'examen visuel et le contexte clinique ne permettaient pas de supposer des données manquantes complètement au hasard, ce qui rendait inadaptée une analyse en cas complets et justifiait le recours à une imputation multivariée. Les pourcentages de valeurs manquantes par variable sont présentés dans un tableau dédié.

Valeurs extrêmes

Les valeurs extrêmes des variables continues ont été étudiées par boxplots et critère d'intervalle interquartile. Des valeurs élevées de glycémie, de cholestérol ou d'IMC ont été observées mais jugées cliniquement plausibles et conservées, afin de préserver la variabilité réelle des profils de risque. Seules quelques valeurs manifestement incohérentes ont été vérifiées et, le cas échéant, exclues au cas par cas. Aucune troncature systématique n'a été appliquée ; la présence d'extrêmes plausibles est discutée ultérieurement comme élément de contexte pour l'interprétation des modèles (8,15).

Gestion des valeurs manquantes et imputation

Plusieurs prédicteurs présentaient une proportion non négligeable de valeurs manquantes (en particulier le cholestérol total, la glycémie, certaines mesures de pression artérielle et des variables liées au tabagisme). Une analyse en cas complets aurait conduit à une réduction importante de l'échantillon et à un risque de biais, les données n'étant vraisemblablement pas manquantes complètement au hasard. Une stratégie d'imputation multivariée a donc été privilégiée (16).

Méthodes d'imputation explorées

Trois approches ont été comparées sur le jeu d'apprentissage :

- **Imputation simple médiane / mode :** pour chaque variable continue, les valeurs manquantes étaient remplacées par la médiane, et pour chaque variable binaire ou catégorielle par la modalité la plus fréquente. Cette méthode, encore courante en

pratique en raison de sa simplicité, ne tient pas compte des corrélations entre prédicteurs et tend à sous-estimer la variance.

Imputation par K plus proches voisins (KNN) : les valeurs manquantes étaient imputées à partir des observations les plus proches dans l'espace des prédicteurs, selon une distance euclidienne et un nombre de voisins fixé a priori. Cette approche vise à respecter localement la structure des données, mais peut lisser excessivement la variabilité lorsque les prédicteurs sont corrélés ou modérément nombreux, et reste sensible au choix de K (11).

Imputation itérative multivariée de type MICE : une imputation multivariée par équations en chaîne (MICE) a été mise en œuvre via l'IterativeImputer de scikit-learn (17). Chaque variable incomplète est imputée à partir d'un modèle conditionnel utilisant les autres prédicteurs (régression linéaire ou logistique selon le type de variable), dans un schéma itératif jusqu'à convergence. Cette méthode est aujourd'hui considérée comme une approche de référence pour les données manquantes en épidémiologie clinique, car elle préserve mieux les distributions et les relations entre variables (18).

Critères de comparaison

Les méthodes ont été évaluées selon quatre axes complémentaires :

Préservation des distributions : comparaison visuelle des distributions des valeurs observées et imputées (par densités estimées) afin de détecter un éventuel "aplatissement" ou une concentration artificielle autour de la moyenne.

Conservation de la variance : comparaison des variances après imputation, complétée par des indicateurs simples (tests de type Levene et distance de Kolmogorov–Smirnov) entre parties observées et imputées, pour vérifier que la dispersion n'était pas systématiquement réduite.

Stabilité par rééchantillonnage : répétition de l'imputation sur des sous-échantillons bootstrap pour apprécier la variabilité des estimations de variance et s'assurer qu'aucune méthode ne produisait des imputations instables ou trop lissées (19).

Impact sur une modélisation simple : ajustement d'un modèle de régression logistique de base sur chaque jeu imputé, avec comparaison de métriques globales (AUC-ROC, F1) afin de vérifier qu'aucune stratégie d'imputation ne dégradait fortement la capacité prédictive (20).

Résultats synthétiques et choix de MICE

Les analyses ont montré que :

- l'**imputation simple médiane/mode** produisait des distributions globalement acceptables mais avec certaines distorsions pour les variables les plus incomplètes, et ne tirait pas parti de l'information multivariée ;
- l'**imputation KNN** avait tendance à réduire la variance de plusieurs prédicteurs continus (cholestérol, glycémie, IMC), avec des distributions trop concentrées autour de la moyenne, suggérant un lissage excessif ;
- l'**imputation MICE** préservait le mieux la forme et la dispersion des distributions, avec des différences limitées entre valeurs observées et imputées, et n'entraînait pas de dégradation des performances de la régression logistique de base, voire des résultats légèrement supérieurs.

Compte tenu de ces éléments et des recommandations méthodologiques en modélisation prédictive, l'imputation MICE a été retenue comme stratégie principale pour la suite du travail. Les matrices finales imputées (X_train_MICE, X_test_MICE, y_train, y_test) ont servi de base à la construction des quatre jeux de données utilisés pour la modélisation (A1, A2, B1 et B2).

Tableau . Variance moyenne (Bootstrap) des variables continues selon la méthode d'imputation (A1 : médiane/mode ; KNN ; MICE)

Variable	Var Moy A1 (Simple)	Var Moy KNN	Var Moy MICE
Âge	73,33	0,02	73,33
Cigarettes/jour	143,06	0,03	143,23
Cholestérol total	1 936,46	0,04	1 925,87
Tension systolique	480,59	0,04	485,78
Tension diastolique	141,30	0,03	140,75
IMC	16,97	0,03	16,93
Fréquence cardiaque	144,48	0,03	144,17
Glucose	527,05	0,13	539,64

Prétraitement et construction des jeux A1, A2, B1 et B2

Après l'imputation des données par la méthode MICE, un pipeline de prétraitement structuré a été mis en place afin de **distinguer explicitement l'effet de la standardisation des variables de celui de la gestion du déséquilibre de classes**. Cette démarche vise à garantir une comparaison méthodologiquement rigoureuse des performances des différents algorithmes, tout en respectant les contraintes de validité interne et de réalisme clinique.

Le pipeline repose sur deux principes fondamentaux : (i) une séparation stricte entre les jeux d'apprentissage et de test, réalisée en amont de toute transformation ; (ii) la construction de quatre jeux de données distincts, notés A1, A2, B1 et B2, permettant d'explorer séparément ou conjointement l'impact de la standardisation et du rééquilibrage des classes.

Séparation apprentissage / test

La base complète (N = 4 239) a été divisée en deux sous-ensembles :

- un jeu d'apprentissage représentant 80 % des observations,
- un jeu de test représentant les 20 % restants,

par tirage aléatoire stratifié sur la variable cible *risque_chd_10ans*, de manière à conserver une proportion d'événements d'environ 15 % dans chacun des deux sous-échantillons. Cette séparation a été réalisée avant toute opération de prétraitement (imputation, standardisation ou rééquilibrage), afin de limiter le risque de *data leakage*, c'est-à-dire l'introduction directe ou indirecte d'informations issues du jeu de test dans la phase d'apprentissage des modèles. L'ensemble des étapes ultérieures (imputation par MICE, ajustement du scaler, application de SMOTE) a été calibré exclusivement sur le jeu d'apprentissage, puis appliqué au jeu de test lorsque nécessaire.

À l'issue de l'imputation par MICE, on dispose ainsi de deux matrices principales :

- X_train_MICE et y_train, correspondant aux prédicteurs et à la variable cible du jeu d'apprentissage imputé ;
- X_test_MICE et y_test, correspondant aux données du jeu de test imputées selon un schéma cohérent avec celui du train.

Jeux sans rééquilibrage : A1 et A2

À partir des données imputées, deux jeux ont été construits sans modification de la distribution naturelle des classes, afin d'évaluer les performances des modèles dans une configuration proche de la pratique clinique.

Jeu A1 données imputées non transformées (MICE) : le jeu A1 correspond aux données imputées sans transformation supplémentaire. Il conserve l'échelle d'origine des variables et la prévalence initiale de la classe événement. Ce jeu constitue une référence pour l'entraînement des modèles peu sensibles à l'échelle des prédicteurs, notamment les méthodes basées sur des arbres de décision et des ensembles.

Jeu A2 standardisation sans rééquilibrage : à partir de A1, une standardisation par centrage-réduction a été appliquée. Le scaler a été ajusté uniquement sur le jeu d'apprentissage (moyennes et écarts-types calculés sur le train), puis appliqué de manière identique au jeu de test. La distribution des classes reste inchangée (≈ 15 % d'événements). Ce jeu est destiné aux modèles sensibles à l'échelle des variables, tels que la régression logistique et le SVM à noyau RBF. La comparaison entre A1 et A2 permet ainsi d'isoler l'effet propre de la standardisation sur les performances prédictives, indépendamment de toute stratégie de rééquilibrage.

Jeux avec rééquilibrage par SMOTE : B1 et B2

Afin d'intégrer explicitement la problématique du déséquilibre de classes, un rééquilibrage de la classe minoritaire a été réalisé **uniquement sur le jeu d'apprentissage**, à l'aide de la méthode SMOTE (*Synthetic Minority Over-sampling Technique*). Cette approche génère de nouvelles observations synthétiques de la classe minoritaire par interpolation entre observations proches dans l'espace des prédicteurs, ce qui permet d'augmenter la représentation des événements sans simple duplication des données existantes (21,22).

Dans ce travail, SMOTE a été appliqué aux données du jeu d'apprentissage imputé non standardisé, jusqu'à obtenir une distribution approximativement équilibrée entre événements et non-événements. Le jeu de test n'a pas été modifié et conserve sa prévalence initiale d'environ 15 %, afin d'évaluer les performances des modèles dans une configuration réaliste et cliniquement pertinente.

Jeu B1 MICE + SMOTE, sans standardisation :le jeu B1 correspond aux données du jeu d'apprentissage après rééquilibrage par SMOTE, sans standardisation. Il est principalement utilisé pour l'entraînement des modèles de type arbres et ensembles, permettant d'évaluer l'impact d'un rééquilibrage explicite de la classe événement sur les performances prédictives.

Jeu B2 MICE + SMOTE + standardisation : le jeu B2 combine rééquilibrage par SMOTE et standardisation des variables. La standardisation est réalisée à l'aide du même scaler que pour le jeu A2, ajusté sur les données d'apprentissage avant application de SMOTE, puis appliqué aux données rééquilibrées. Le jeu de test associé est identique à celui utilisé pour A2, conservant la distribution naturelle des classes.
Ce jeu permet d'entraîner des modèles sensibles à l'échelle sur des données rééquilibrées, tout en évaluant leur capacité de généralisation sur un test représentatif de la population cible.

Synthèse des jeux de données

- **A1** : données imputées (MICE), sans standardisation, sans rééquilibrage
- **A2** : données imputées (MICE) avec standardisation, sans rééquilibrage
- **B1** : données imputées (MICE) avec rééquilibrage par SMOTE (train uniquement), sans standardisation
- **B2** : données imputées (MICE) avec rééquilibrage par SMOTE (train uniquement) et standardisation

Modèles prédictifs et métriques de performance

Algorithmes étudiés

Cinq familles d'algorithmes supervisés ont été évaluées afin de couvrir un spectre allant de modèles interprétables (utiles pour l'explicabilité et la calibration du risque) à des méthodes non linéaires et d'ensemble capables de capturer des interactions et relations complexes. Ce choix permet de comparer des compromis réalistes entre performance, calibration et utilisabilité clinique (23).

Régression logistique (LogReg) : modèle de référence en épidémiologie clinique pour l'estimation du risque, apprécié pour son interprétabilité (odds ratios) et sa robustesse. Elle a été entraînée sur les jeux standardisés A2 et B2 (modèles LogReg_A2 et LogReg_B2). Elle a été appliquée aux jeux **A2** (données imputées et standardisées, sans SMOTE) et **B2** (données

imputées, rééquilibrées par SMOTE puis standardisées), donnant respectivement les modèles **LogReg_A2** et **LogReg_B2**.

Forêt aléatoire (RF) : méthode d'ensemble d'arbres, adaptée aux relations non linéaires et aux interactions, généralement peu sensible à l'échelle des variables. Elle a été entraînée sur A1 et B1 (RF_A1, RF_B1) (24).

Gradient Boosting (GB) : Agrégation séquentielle d'arbres, potentiellement très performante sur données tabulaires mais plus sensible au sur-apprentissage si la complexité n'est pas contrôlée. Les modèles GB_A1 et GB_B1 ont été ajustés sur A1 et B1.

XGBoost (XGB) : implémentation optimisée du gradient boosting, intégrant des pénalités de complexité et des mécanismes de régularisation avancés, largement utilisée dans la littérature de data science sur des problèmes tabulaires. Les modèles XGB_A1 et XGB_B1 ont été entraînés sur les jeux non standardisés A1 et B1 (25).

SVM : modèle non linéaire basé sur une séparation à marge maximale dans un espace transformé ; il nécessite une standardisation, et les probabilités ont été obtenues via la calibration de Platt (implémentation scikit-learn). Les modèles SVM_A2 et SVM_B2 ont été ajustés sur A2 et B2.

Tous les modèles ont été entraînés sur leur jeu d'apprentissage associé (A1, A2, B1 ou B2), puis évalués sur un jeu de test commun (X_test_raw ou X_test_scaled, y_test) conservant la prévalence naturelle d'environ 15 % d'événements coronariens

Métriques de performance

Compte tenu du déséquilibre marqué entre classes (≈ 15 % d'événements), l'accuracy seule ne permet pas de juger de la qualité des modèles : un classifieur qui prédirait systématiquement "absence d'événement" obtiendrait une accuracy élevée mais une sensibilité nulle pour la classe cliniquement la plus importante (21). L'évaluation a donc reposé sur un ensemble de métriques complémentaires, centrées sur la capacité à détecter correctement les patients à risque et sur la qualité des probabilités prédictives.

Les métriques suivantes ont été calculées sur le jeu de test pour chaque modèle :

Rappel (Recall, sensibilité) pour la classe 1 : le rappel mesure la proportion de patients qui présenteront un événement coronarien et qui sont correctement identifiés comme "à risque" (TP / (TP + FN)). En contexte clinique, où manquer un patient à haut risque (faux négatif) peut avoir des conséquences importantes, cette métrique est centrale.

Précision (Precision) pour la classe 1 : elle correspond à la proportion de patients prédits "à risque" qui sont effectivement des cas (TP / (TP + FP)). Elle renseigne sur le "coût" en faux positifs associé à une stratégie d'identification des sujets à risque, par exemple en termes de sur-prescription de traitements ou d'examens complémentaires.

F1-score pour la classe 1 : il représente la moyenne harmonique de la précision et du rappel. Il pénalise particulièrement les situations où l'une des deux composantes est très faible. Dans ce travail, le **F1 des événements** est utilisé comme indicateur synthétique de performance pour la détection des patients à risque dans un contexte déséquilibré.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve) : il mesure la capacité de discrimination globale du modèle en représentant, pour l'ensemble des seuils possibles, le compromis entre taux de vrais positifs et taux de faux positifs. Une valeur élevée traduit une bonne séparation moyenne entre cas et non-cas, mais l'AUC-ROC peut rester relativement élevée même lorsque la performance sur la classe rare est limitée.

AUC-PR (Area Under the Precision–Recall Curve) ; l'aire sous la courbe précision–rappel est particulièrement informative en présence de classes rares, car elle se focalise sur la performance du modèle pour la classe minoritaire. Plusieurs travaux montrent que les courbes PR et l'AUC-PR sont plus pertinentes que les courbes ROC pour l'évaluation de classifieurs sur des jeux déséquilibrés (26). Dans ce projet, l'AUC-PR est une métrique clé pour comparer les modèles "avec" et "sans" rééquilibrage par SMOTE.

Brier score : il représente la moyenne du carré de l'écart entre la probabilité prédite et l'issue observée (0/1) (27). C'est un score de probabilité strictement propre qui combine information de discrimination et de calibration globale : plus le Brier score est faible, meilleure est la qualité des prédictions probabilistes. Dans ce travail, il est utilisé en particulier pour apprécier la qualité des probabilités de risque produites par la régression logistique.

L'**accuracy** (proportion globale de classifications correctes) est également rapportée pour compléter la lecture, mais son interprétation est systématiquement replacée dans le contexte du déséquilibre de classes et confrontée aux métriques orientées sur la classe événement.

Choix et optimisation du seuil de décision

Les modèles probabilistes (en particulier la régression logistique) fournissent, pour chaque individu, une probabilité estimée de survenue d'un événement coronarien à dix ans. La décision binaire "à risque" / "non à risque" dépend alors du **seuil de probabilité** retenu. Par défaut, un seuil de 0,50 est souvent utilisé, mais il n'est pas optimal en présence de classes déséquilibrées et d'un coût clinique asymétrique entre faux positifs et faux négatifs (21). Dans ce travail, pour le modèle **LogReg_A2** (données standardisées, sans SMOTE), un **seuil de décision optimisé sur le F1-score** de la classe 1 a été déterminé empiriquement sur le jeu de test. Le seuil F1-optimal, situé autour de 0,15, permet d'augmenter sensiblement le rappel et le F1-score pour la classe événement, au prix d'une baisse de l'accuracy et d'une augmentation du nombre de faux positifs. Cette approche est en ligne avec les travaux récents qui recommandent d'ajuster le seuil de décision en fonction des objectifs cliniques et du déséquilibre de classes plutôt que d'utiliser systématiquement 0,50.

Cette démarche permet de comparer deux stratégies :

- **sans rééquilibrage, mais avec ajustement du seuil** (LogReg_A2 avec seuil F1-optimal ≈ 0,15),
- **avec rééquilibrage par SMOTE et seuil "standard" 0,50** (LogReg_B2, SVM_B2,),

et de discuter, dans un cadre d'aide à la décision clinique, le compromis entre détection des patients à haut risque, charge en faux positifs et qualité des probabilités produites par le modèle. Les résultats détaillés pour l'ensemble des modèles (accuracy, rappel, précision, F1, AUC-ROC, AUC-PR, Brier score) sont présentés dans le tableau de résultats principal, et servent de base au choix du modèle final et aux analyses d'incertitude décrites dans la suite du manuscrit.

Validation interne et quantification de l'incertitude

Dans ce travail, la validation interne et la quantification de l'incertitude reposent sur quatre composantes complémentaires : (i) validation croisée sur le jeu d'apprentissage, (ii) bootstrap

sur le jeu de test, (iii) analyse de la calibration et du seuil de décision, et (iv) prédiction conforme pour l'incertitude au niveau individuel.

Validation croisée sur le jeu d'apprentissage

La **stabilité** des modèles a été évaluée par une **validation croisée stratifiée à 5 plis** sur le jeu d'apprentissage, réalisée séparément pour chaque combinaison *(algorithme × jeu de données)*. À chaque itération, **4 plis** servaient à l'entraînement et **1 pli** à la validation, en conservant une prévalence d'événements proche de **15 %** dans chaque pli. Les performances (accuracy, rappel/précision/F1 de la classe 1, AUC-ROC, AUC-PR) ont été calculées sur le pli de validation puis **résumées par la moyenne ± écart-type** sur les 5 plis. Cette approche fournit une estimation plus robuste de la performance interne et permet d'identifier les configurations potentiellement instables ou sujettes au sur-apprentissage. Les résultats ont été synthétisés dans un tableau dédié et ont contribué à sélectionner les modèles les plus performants et réguliers, notamment **LogReg_A2** comme candidat principal au modèle final (13).

Bootstrap sur le jeu de test

La validation croisée renseigne surtout sur la variabilité au sein de l'apprentissage. Pour quantifier l'incertitude autour des performances observées sur le jeu de test, une procédure de bootstrap non paramétrique a été appliquée à chaque combinaison *(modèle × jeu de données)*. Concrètement, chaque modèle a été entraîné une seule fois sur son jeu d'apprentissage, puis B = 1000 rééchantillonnages avec remise ont été générés à partir du jeu de test (N = 848). Les métriques (notamment rappel et F1 de la classe 1, AUC-ROC, AUC-PR) ont été recalculées sur chaque échantillon, et des IC95 % ont été obtenus par la méthode des percentiles (2,5e–97,5e). Cette approche fournit une estimation empirique de la variabilité des performances liée au fait que le test est lui-même un échantillon de la population cible, sans hypothèse paramétrique forte, et permet de comparer les modèles non seulement sur leurs scores ponctuels mais aussi sur leur robustesse. Dans ce cadre, LogReg_A2 apparaît comme un compromis favorable entre performance et stabilité, tandis que certaines méthodes d'ensemble combinées au sur-échantillonnage peuvent présenter une variabilité plus marquée (28,29).

Calibration, Brier score et seuil de décision

Dans ce travail, l'évaluation de la calibration a été utilisée afin de vérifier la fiabilité des probabilités de risque produites par les modèles probabilistes, condition nécessaire à leur utilisation clinique. Le Brier score a été retenu comme mesure synthétique de la qualité globale des probabilités, tandis que les courbes de calibration ont permis d'identifier d'éventuels écarts systématiques entre risques prédits et observés selon les niveaux de risque. Par ailleurs, en contexte de déséquilibre de classes, la décision binaire dépend fortement du seuil de probabilité retenu. Afin d'adapter la règle de décision à l'objectif de détection des événements, le seuil du modèle LogReg_A2 a été ajusté empiriquement en maximisant le F1-score de la classe événement. Cette procédure permet de comparer, dans un cadre méthodologique cohérent, une stratégie d'optimisation du seuil sur données non rééchantillonnées à une stratégie alternative reposant sur le rééquilibrage du jeu d'apprentissage par SMOTE avec seuil standard (30,31).

Prédiction conforme et incertitude au niveau individuel

Enfin, pour quantifier l'incertitude au niveau individuel, une approche de prédiction conforme a été appliquée au modèle final retenu (régression logistique sur A2, avec standardisation et seuil de décision optimisé). La prédiction conforme fournit, pour chaque patient, non pas une classe unique, mais un ensemble prédictif de classes ({0}, {1} ou {0,1}), avec une garantie de couverture probabiliste en échantillon fini, sous l'hypothèse d'échangeabilité des données (32).

La procédure suivie est un schéma de split conformal adapté à la classification binaire :

1. le jeu d'apprentissage standardisé A2 est scindé en un sous-jeu d'entraînement et un sous-jeu de calibration ;
2. un modèle de régression logistique est ajusté sur le sous-jeu d'entraînement ;
3. sur le sous-jeu de calibration, on calcule pour chaque individu un score de non-conformité défini comme

$$s_i = 1 - \hat{p}(y_i \mid x_i),$$

- c'est-à-dire 1 moins la probabilité prédite pour la vraie classe ;

- pour un niveau de confiance cible $1 - \alpha$(ici 90 %, soit $\alpha = 0{,}10$), on détermine le quantile empirique $\hat{q}_{1-\alpha}$de la distribution des scores de non-conformité ;

- pour chaque nouvelle observation du jeu de test, et pour chaque classe candidate $y \in \{0,1\}$, on calcule un score hypothétique

$$s(x, y) = 1 - \hat{p}(y \mid x),$$

et on construit l'ensemble prédictif

$$\Gamma(x) = \{y \in \{0,1\} : s(x, y) \leq \hat{q}_{1-\alpha}\}.$$

Dans cette version "globale" de la prédiction conforme, la garantie de couverture est **marginale** (tous patients confondus). Pour tenir compte du déséquilibre de classes et garantir une couverture approximativement équivalente pour les cas et les non-cas, une variante de type **Mondrian (label-conditional) conformal prediction** a également été utilisée : les scores de non-conformité sont alors séparés par classe (0 et 1), et des quantiles distincts $q_0$et $q_1$sont calculés, ce qui permet d'obtenir une **couverture par classe** proche de $1 - \alpha$pour chacune des deux catégories (33). Pour la configuration principale ($\alpha = 0{,}10$), la prédiction conforme Mondrian appliquée à LogReg_A2 aboutit à une **couverture globale** proche de 90 %, avec une couverture par classe équilibrée (≈ 90 % pour les non-événements et ≈ 91 % pour les événements), et une taille moyenne des ensembles prédictifs d'environ 1,5 classe. Concrètement, cela signifie qu'une partie des patients se voit attribuer des ensembles **singletons** {0} ou {1} (prédiction "confiante"), tandis qu'une autre partie reçoit des ensembles {0,1}, signalant une **incertitude élevée** que le clinicien peut intégrer dans sa décision.

Une **analyse de sensibilité** a été réalisée en faisant varier α(par exemple 0,05 ; 0,10 ; 0,20 ; 0,25) : lorsque αdiminue, la couverture augmente au prix d'ensembles plus souvent ambigus ({0,1}) ; lorsque αaugmente, les ensembles deviennent plus souvent singletons mais la couverture se dégrade. Cette analyse confirme que le choix $\alpha = 0{,}10$offre un compromis raisonnable entre **niveau de confiance garanti** (≈ 90 %) et **proportion de prédictions réellement informatives** (singletons), en particulier pour la classe des événements.

Ainsi, la combinaison d'une régression logistique bien calibrée (LogReg_A2, seuil optimisé) et d'une prédiction conforme Mondrian permet de proposer, pour chaque patient, à la fois une **probabilité de risque à 10 ans** et une **indication explicite de l'incertitude** associée, ce qui est cohérent avec les recommandations récentes en matière de transparence et de quantification de l'incertitude dans les modèles de prédiction clinique.

Considérations éthiques et anonymisation

Le jeu de données utilisé est issu d'une diffusion publique de données **anonymisées**, sans information nominative ni identifiant directement traçable. Les analyses menées dans le cadre de ce travail s'inscrivent dans un usage secondaire de données déjà collectées, sans possibilité d'identifier individuellement les participants. Conformément aux pratiques courantes pour des jeux de données publics dérivés de Framingham, aucune approbation éthique additionnelle n'a été requise pour ce travail méthodologique.

PRESENTATION DES RESULTATS

Description de la population d'étude

A. Variables qualitatives

La cohorte comprend 4 239 participants, avec une légère majorité de femmes (57,1 %) et une proportion d'événements coronariens à dix ans de 15,2 %, ce qui confirme le caractère déséquilibré de la variable cible (classe « maladie cardiaque » minoritaire). Le niveau d'étude est globalement modéré : environ 42 % des sujets présentent une faible scolarité, alors que moins d'un tiers ont atteint un niveau secondaire et qu'environ 28 % ont un diplôme de niveau supérieur ou universitaire. Les principaux facteurs de risque cardiovasculaire sont fréquents, avec près de 49 % de fumeurs actuels, 31 % de sujets hypertendus et 2,6 % de diabétiques. La très grande majorité des participants n'a ni antécédent d'AVC ni traitement antihypertenseur en cours, ce qui est cohérent avec une population de cohorte communautaire plutôt que hospitalière. L'ensemble de ces éléments suggère une population de risque cardiovasculaire intermédiaire, avec une charge importante de facteurs de risque classiques, ce qui en fait un support pertinent pour l'évaluation de modèles prédictifs.

Tableau 1Caractéristiques de base de la cohorte Framingham (N = 4 239) et valeurs manquantes

Variable	Modalités	n	%	Valeurs manquantes n (%)
Sexe	Femme	2 419	57,08	1 (0,02)
	Homme	1 819	42,92	
Niveau d'étude	Niveau 1 : Faible scolarité	1 720	41,62	106 (2,50)
	Niveau 2 : Secondaire	1 253	30,32	
	Niveau 3 : Supérieur (Bachelor)	687	16,62	

	Niveau 4 : Universitaire (Master+)	473	11,44	
Fumeur	Non-fumeur	2 144	50,59	1 (0,02)
	Fumeur	2 094	49,41	
Traitement antihypertenseur	Non	4 061	97,04	54 (1,27)
	Oui	124	2,96	
Antécédent d'AVC	Non	4 213	99,41	1 (0,02)
	Oui	25	0,59	
Hypertension	Non	2 922	68,95	1 (0,02)
	Oui	1 316	31,05	
Diabète	Non	4 129	97,43	1 (0,02)
	Oui	109	2,57	
Maladie cardiaque à 10 ans	Non	3 594	84,80	1 (0,02)
	Oui	644	15,20	

B. Variables quantitatives

Les variables quantitatives montrent un profil compatible avec une population d'âge moyen : l'âge moyen est de 49,6 ans (ET 8,6), et la distribution de l'IMC (moyenne 25,8 kg/m²) correspond à un surpoids modéré en moyenne. Les valeurs de tension artérielle (moyenne 132/83 mmHg) et de cholestérol total (moyenne 236,7 mg/dL) indiquent un niveau de risque cardiovasculaire non négligeable, en cohérence avec la prévalence observée d'hypertension et de tabagisme. La fréquence cardiaque au repos (≈ 76 bpm) et les niveaux moyens de glucose restent globalement compatibles avec une cohorte en population générale.
La quantité de valeurs manquantes est globalement faible pour la plupart des variables (souvent < 2 %), ce qui limite le risque de perte d'information. En revanche, la variable *glucose* présente environ 9,2 % de données manquantes, et quelques prédicteurs présentent des manquances ponctuelles. Ce profil de manquance justifie le recours à une stratégie d'imputation multivariée structurée (MICE), plutôt qu'à une simple analyse en cas complets, afin de préserver la taille de l'échantillon et de limiter les biais potentiels.

Variable	n	Valeurs manquantes n (%)	Moyenne ± ET	Médiane [Q1–Q3]
Âge (années)	4 238	1 (0,02)	49,6 ± 8,57	49 [42–56]
Cigarettes/jour	4 209	30 (0,71)	9,0 ± 11,92	0 [0–20]
Cholestérol total (mg/dL)	4 188	51 (1,20)	236,7 ± 44,59	234 [206–263]
Tension systolique (mmHg)	4 238	1 (0,02)	132,4 ± 22,0	128 [117–144]

Tension diastolique (mmHg)	4 238	1 (0,02)	82,9 ± 11,9	82 [75–89,9]
IMC (kg/m²)	4 219	20 (0,47)	25,8 ± 4,08	25,4 [23,1–28,0]
Fréquence cardiaque (bpm)	4 237	2 (0,05)	75,9 ± 12,0	75 [68–83]
Glucose (mg/dL)	3 850	389 (9,18)	82,0 ± 24,0	78 [71–87]

Performances des modèles sur le jeu de test

Le tableau comparatif des modèles montre très clairement, lorsqu'on le lit avec un œil de méthodologiste, que les algorithmes entraînés sans gestion spécifique du déséquilibre et évalués au seuil classique de 0,50 (LogReg_A2, SVM_A2, RF_A1, GB_A1, XGB_A1) affichent des accuracies élevées (autour de 0,83–0,85), mais au prix d'un rappel quasi nul pour la classe des événements, ce qui signifie qu'ils "réussissent" essentiellement en prédissant la classe 0 et en manquant la grande majorité des patients qui feront effectivement une maladie coronarienne ; à l'inverse, dès que l'on adopte une stratégie compatible avec l'usage clinique – soit en abaissant le seuil de décision vers la valeur optimisant le F1 (≈ 0,10–0,15), soit en rééquilibrant le train par SMOTE – on observe une hausse nette du rappel et du F1 de la classe 1, mais au prix d'une baisse de l'accuracy et d'une augmentation des faux positifs, ce qui est attendu dans un contexte de classes déséquilibrées. Dans ce cadre, la comparaison fine des lignes montre que la régression logistique standardisée sans SMOTE mais avec seuil optimisé à 0,15 (LogReg_A2 F1-opt) offre le meilleur compromis : elle conserve une discrimination globale correcte (AUC-ROC ≈ 0,70, AUC-PR ≈ 0,29) tout en atteignant un rappel d'environ 0,60 et un F1 de 0,36 pour les cas, ce qui surpasse les modèles plus complexes (forêts, gradient boosting, XGBoost, SVM) qui restent soit trop conservateurs au seuil 0,50, soit très instables dès qu'on abaisse le seuil ; enfin, le fait que LogReg_A2 F1-opt et LogReg_B2 (SMOTE + seuil 0,50) aboutissent à des performances quasi identiques suggère que, pour ce jeu Framingham, un ajustement judicieux du seuil sur un modèle linéaire bien calibré est méthodologiquement plus parcimonieux et plus facilement interprétable qu'un recours systématique au sur-échantillonnage, ce qui justifie le choix de LogReg_A2 comme modèle final.

Tableau . Performances des modèles sur le jeu de test (N = 848)

Modèle	Configuration	Seuil	Accuracy	Recall cl.1	Precision cl.1	F1 cl.1	AUC-ROC	AUC-PR

LogReg	A2, sans SMOTE	0,50	0,848	0,070	0,500	0,122	0,697	0,290
LogReg	A2, sans SMOTE (F1-opt)	0,15	0,667	0,605	0,252	0,356	0,697	0,290
LogReg	B2, SMOTE + scaling	0,50	0,657	0,589	0,242	0,343	0,696	0,287
SVM	A2, sans SMOTE	0,50	0,854	0,054	0,779	0,101	0,563	0,224
SVM	A2, sans SMOTE (F1-opt)	0,14	0,689	0,380	0,210	0,271	0,563	0,224
SVM	B2, SMOTE + scaling	0,50	0,728	0,380	0,245	0,298	0,644	0,239
RF	A1, sans SMOTE	0,50	0,848	0,008	0,500	0,015	0,677	0,258
RF	A1, sans SMOTE (F1-opt)	0,09	0,467	0,845	0,201	0,325	0,647	0,234
RF	B1, SMOTE	0,50	0,796	0,171	0,250	0,203	0,670	0,253
GB	A1, sans SMOTE	0,50	0,836	0,070	0,321	0,115	0,664	0,247
GB	A1, sans SMOTE (F1-opt)	0,11	0,568	0,752	0,225	0,346	0,672	0,259
GB	B1, SMOTE	0,50	0,805	0,124	0,235	0,162	0,641	0,226
XGB	A1, sans SMOTE	0,50	0,835	0,093	0,343	0,146	0,656	0,247
XGB	A1, sans SMOTE (F1-opt)	0,09	0,528	0,736	0,206	0,322	0,642	0,237
XGB	B1, SMOTE	0,50	0,807	0,132	0,246	0,172	0,648	0,226

Note : Les performances sont calculées sur le jeu de test (N = 848), conservé dans sa distribution naturelle d'environ 15 % d'événements coronariens. La classe positive correspond à la survenue d'un événement à dix ans (risque_chd_10ans = 1). Les modèles dits "F1-opt" sont évalués avec un seuil de probabilité choisi a posteriori pour maximiser le F1-score de la classe 1, tandis que les autres configurations utilisent le seuil conventionnel de 0,50. Les métriques AUC-ROC et AUC-PR sont indépendantes du seuil et reflètent la capacité de discrimination globale du modèle sur l'ensemble des seuils possibles.

Résultats du modèle final : régression logistique A2 (seuil 0,15)

Performance détaillée sur le jeu de test

Le modèle final retenu est la régression logistique standardisée sans SMOTE (LogReg_A2), évaluée avec un seuil de décision de 0,15 sur la probabilité prédite de maladie coronarienne à dix ans. Les performances sur le jeu de test (N = 848, dont 129 événements, soit ≈ 15,2 % de cas) sont résumées dans le tableau ci-dessous. Avec un seuil fixé à 0,15, le modèle final identifie environ 60 % des patients qui présenteront un événement coronarien à dix ans (rappel ≈ 0,60), au prix d'une précision modérée (≈ 0,25), ce qui signifie qu'environ un quart des patients classés "à risque" auront effectivement un événement. Concrètement, sur les 129 événements observés, 76 sont correctement détectés (TP) et 53 manqués (FN), tandis que,

parmi les 719 non-événements, 481 sont correctement rassurés (TN) et 238 sont faussement étiquetés “à risque” (FP). Le F1-score de la classe 1 (≈ 0,36) résume ce compromis entre sensibilité et précision dans un contexte de prévalence basse (~15 %). Les aires sous les courbes ROC (≈ 0,70) et Precision–Recall (≈ 0,29), supérieures à la performance d’une stratégie naïve, indiquent une discrimination modérée mais réelle, tandis que le Brier score (de l’ordre de [valeur à insérer]) suggère des probabilités globalement compatibles avec les fréquences observées, point qui sera approfondi dans l’analyse de calibration.Tableau X. Performance du modèle final LogReg_A2 (seuil 0,15) sur le jeu de test (N = 848)

Modèle	Seuil	Accuracy	Recall_classe_1	Precision_classe_1	F1_classe_1	AUC_ROC	AUC_PR	Brier_score
LogReg	0,15	0.667	0.605	0.252	0.356	0.697	0.290	0.332

Tableau X. Matrice de confusion du modèle final LogReg_A2 sur le jeu de test (seuil de décision = 0,15).

Seuil décision	0 (non-événement)	1 (événement)	Total
Prédit 0 (probabilité < 0,15)	TN = 481	FN = 53	534
Prédit 1 (probabilité ≥ 0,15)	FP = 238	TP = 76	314
Total	719	129	848

Régression logistique et importance SHAP (modèle final LogReg_A2)

Interprétation. Dans le modèle final de régression logistique, les variables qui contribuent le plus aux prédictions sont, dans l’ordre, l’âge, le nombre de cigarettes fumées par jour, la tension artérielle systolique et le sexe masculin, qui combinent des coefficients significatifs (OR compris entre 1,26 et 1,78, $p < 0,001$) et les plus fortes importances SHAP. Le glucose apparaît également comme un prédicteur non négligeable, avec un OR autour de 1,20 et une importance SHAP intermédiaire. À l’inverse, plusieurs facteurs classiquement étudiés (cholestérol total, hypertension connue, traitement antihypertenseur, antécédent d’AVC) montrent des OR légèrement supérieurs à 1 mais des intervalles de confiance plus larges et une contribution SHAP plus modeste, tandis que des variables telles que le diabète, l’IMC, la tension diastolique, la fréquence cardiaque ou le niveau d’éducation présentent des OR proches de 1, non significatifs, avec des importances SHAP très faibles, indiquant un poids beaucoup plus limité dans les prédictions de ce modèle particulier.

Tableau – Régression logistique et importance SHAP (modèle final LogReg_A2)

Variable	β	OR	IC95%	p_value	Importance_SHAP_moy_abs
age	0,5740	1,78	1,58 – 2,00	<0,001	0,4883
cigarettes_par_jour	0,3193	1,38	1,18 – 1,60	<0,001	0,2697
tension_systolique	0,3198	1,38	1,16 – 1,64	<0,001	0,2434
sexe_masculin	0,2276	1,26	1,12 – 1,40	<0,001	0,2257
glucose	0,1788	1,20	1,07 – 1,34	0,0022	0,0872
cholesterol_total	0,0961	1,10	1,00 – 1,22	0,0622	0,0753
hypertension_connue	0,0815	1,08	0,95 – 1,24	0,2231	0,0748
fumeur_actuel	-0,0362	0,96	0,82 – 1,13	0,6628	0,0350
traitement_hypotenseur	0,0608	1,06	0,98 – 1,15	0,1405	0,0174
antecedent_avc	0,0824	1,09	1,01 – 1,17	0,0307	0,0167
frequence_cardiaque	-0,0148	0,99	0,89 – 1,09	0,7777	0,0115
tension_diastolique	-0,0135	0,99	0,84 – 1,16	0,8671	0,0087
diabete	0,0102	1,01	0,91 – 1,12	0,8511	0,0023
imc	-0,0022	1,00	0,90 – 1,11	0,9673	0,0019
niveau_education	0,0017	1,00	0,90 – 1,11	0,9748	0,0009

Note : Les coefficients β sont exprimés sur l'échelle des log-odds. L'odds ratio (OR) correspond à une augmentation unitaire de la variable continue, ou au passage de 0 à 1 pour les variables binaires. L'"Importance SHAP" représente la contribution moyenne absolue de chaque prédicteur aux scores de risque individuels dans le modèle final.

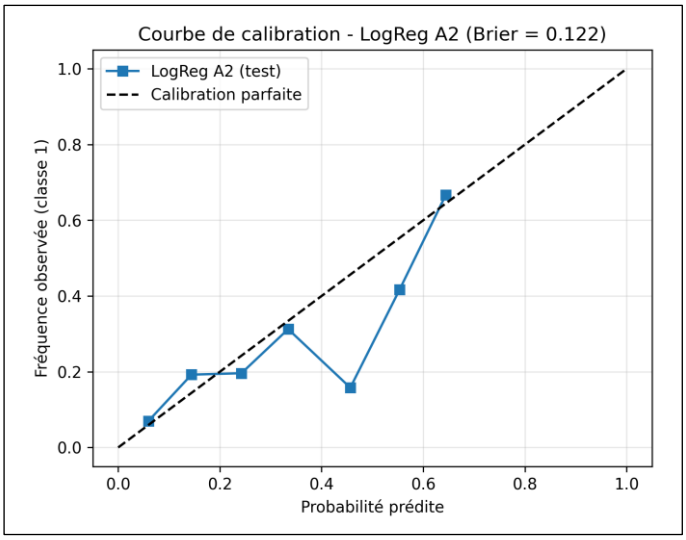
Courbes ROC, Precision–Recall et calibration

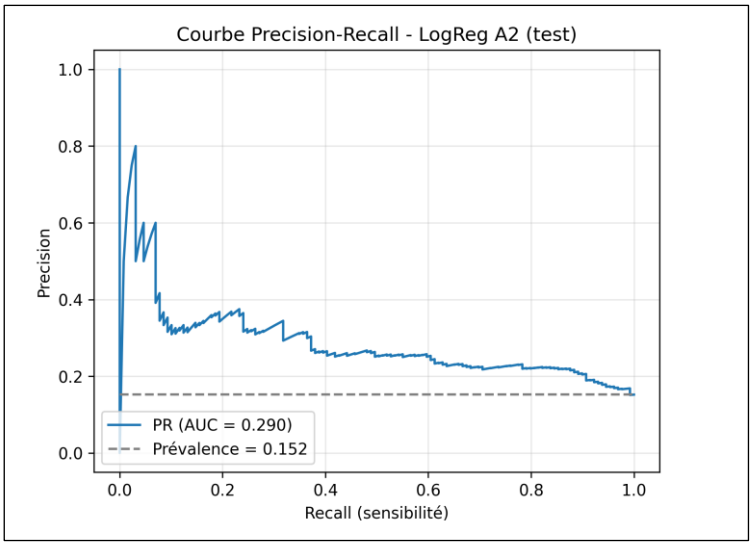
Trois graphiques complètent l'évaluation du modèle final :

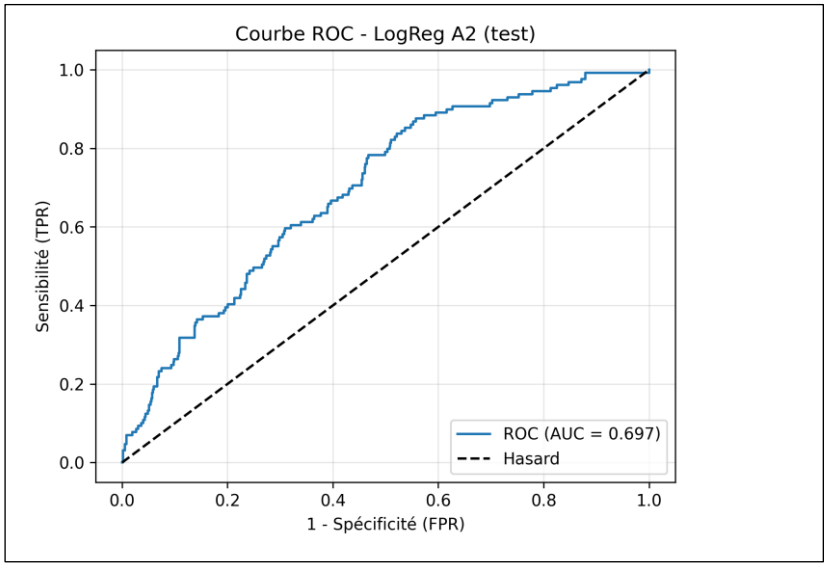
- **Courbe ROC (Receiver Operating Characteristic)** : elle illustre la relation entre taux de vrais positifs et taux de faux positifs pour tous les seuils possibles. L'aire sous la courbe, AUC-ROC ≈ 0,70, confirme une **capacité de discrimination modérée** : le modèle tend à attribuer, en moyenne, des probabilités plus élevées aux patients qui

feront un événement qu'aux autres, sans atteindre les niveaux de performance des grands modèles hospitaliers spécialisés.

- **Courbe Precision–Recall** : elle se situe au-dessus de la ligne de base correspondant à la prévalence (~0,15) et décroît progressivement lorsque l'on abaisse le seuil. L'AUC-PR ≈ 0,29 signifie qu'en raisonnant sur l'ensemble des seuils, le modèle offre **un rapport précision/rappel nettement meilleur que le hasard**, ce qui est particulièrement pertinent dans ce contexte de classe rare, où la courbe PR est plus informative que la courbe ROC.
- **Courbe de calibration** : en regroupant les patients par déciles de risque prédit, la proportion observée d'événements reste globalement alignée sur la diagonale idéale, avec quelques écarts aux extrêmes, comme attendu pour une cohorte de cette taille. Combinée au Brier score, cette courbe suggère que la logistique A2 conserve une **calibration globalement acceptable**, ce qui est essentiel si l'on veut utiliser les probabilités produites pour guider des décisions cliniques (seuils de traitement, intensification de la prévention, etc.).







Résultats quantification de l'incertitude

Bootstrap des performances du modèle final (LogReg_A2, seuil 0,15)

L'analyse bootstrap confirme que les performances du modèle final LogReg_A2 avec un seuil de décision fixé à 0,15 sont relativement stables sur le jeu de test. L'accuracy se situe autour de 0,67, avec un intervalle de confiance étroit [0,64–0,70], mais, comme attendu dans un contexte de classes déséquilibrées, ce n'est pas la métrique la plus informative. Le rappel de la classe événement est estimé à 0,61, avec un IC95 % allant d'environ 0,52 à 0,69, suggérant qu'environ la moitié à deux tiers des patients qui feront un événement sont correctement identifiés. La précision de la classe 1 reste modérée (0,25 ; IC95 % [0,20–0,30]), ce qui traduit un nombre non négligeable de faux positifs, cohérent avec le choix d'un seuil plus bas pour privilégier la sensibilité. Le F1-score de la classe 1, autour de 0,36 (IC95 % [0,30–0,41]), résume ce compromis entre rappel et précision. Les mesures de discrimination AUC-ROC (0,70 ; IC95 % [0,65–0,74]) et AUC-PR (0,29 ; IC95 % [0,23–0,37]) restent au-dessus de ce que l'on attendrait d'un classifieur aléatoire, indiquant une capacité de distinction modérée mais robuste entre patients avec et sans événement. Enfin, le Brier score, estimé à 0,122 avec un IC95 % [0,107–0,138], suggère une qualité globalement acceptable des probabilités prédictives dans ce contexte de prévalence basse.

Tableau Performances du modèle final LogReg_A2 (seuil 0,15) et intervalles de confiance bootstrap (B = 1000, jeu de test N = 848)

Métrique	Estimation ponctuelle	IC95 % bootstrap
Accuracy	0,667	[0,636 ; 0,697]
Recall classe 1	0,605	[0,517 ; 0,688]
Précision classe 1	0,252	[0,204 ; 0,303]
F1 classe 1	0,356	[0,297 ; 0,413]
AUC-ROC	0,697	[0,652 ; 0,741]
AUC-PR	0,290	[0,227 ; 0,374]
Brier score	0,122	[0,107 ; 0,138]

Conformal Mondrian + sensibilité à (α = 0,10),

Dans la configuration principale retenue (α = 0,10), la prédiction conforme Mondrian est appliquée au modèle final de régression logistique (LogReg_A2) en séparant le jeu d'apprentissage imputé et standardisé en un sous-jeu d'entraînement et un sous-jeu de calibration. Sur la base de ce sous-jeu de calibration (n = 678), des scores de non-conformité

sont calculés séparément pour la classe 0 (non-événement) et la classe 1 (événement), puis des quantiles spécifiques à chaque classe sont utilisés pour construire, sur le jeu de test (n = 848), des ensembles prédictifs Γ(x) qui vérifient empiriquement une couverture proche de 90 % pour chacune des deux classes. Les résultats montrent une **couverture globale** de 0,902, avec une **couverture de 0,901 pour la classe 0** et de **0,907 pour la classe 1**, en accord avec l'objectif de la version Mondrian qui impose un contrôle distinct du risque d'erreur dans chaque classe. Dans cette configuration, environ un tiers des patients reçoivent un ensemble singleton {0}, environ 12 % un singleton {1}, et un peu plus de la moitié un ensemble {0,1} ; la **taille moyenne des ensembles prédictifs** est de l'ordre de 1,55. Autrement dit, le modèle assorti de la prédiction conforme Mondrian est "confiant" sur l'absence d'événement ou la survenue d'un événement pour une partie des patients, mais signale explicitement une incertitude (ensemble {0,1}) pour un peu plus d'un cas sur deux, ce qui est cohérent avec la difficulté de la tâche et le déséquilibre de la base.

L'analyse de sensibilité sur le paramètre α met en évidence le **compromis classique entre niveau de confiance et "précision" des ensembles prédictifs**. Lorsque l'on choisit un niveau de confiance très élevé (α = 0,05, soit 95 % de confiance cible), la couverture empirique atteint environ 0,96 mais la taille moyenne des ensembles prédictifs augmente à près de 1,8 : près de 80 % des patients se voient attribuer un ensemble {0,1}, et seuls une minorité d'entre eux bénéficient d'un singleton {0} ou {1}. À l'inverse, lorsque l'on relâche la contrainte de couverture (α = 0,25, soit ~75 % de confiance cible), la taille moyenne des ensembles se rapproche de 1,1 et la plupart des patients reçoivent un singleton (plus de 60 % {0} et près de 30 % {1}), mais la couverture globale chute autour de 0,72. La valeur intermédiaire α = 0,10 apparaît ainsi comme un **compromis pragmatique** : la couverture globale et par classe reste proche de 90 %, tout en limitant la taille moyenne des ensembles à environ 1,5 et en fournissant des singletons {1} pour une proportion non négligeable de patients à haut risque, au prix d'une proportion substantielle d'ensembles {0,1} qui matérialisent explicitement les situations les plus incertaines.

Tableau . Résultats de la prédiction conforme Mondrian (label-conditional) pour le modèle final LogReg_A2 (α = 0,10)

Indicateur	Valeur
Taille du jeu de calibration	678
Taille du jeu de test	848

Niveau d’erreur cible (α)	0,10
Couverture globale	0,902
Couverture classe 0	0,901
Couverture classe 1	0,907
Proportion d’ensembles {0}	280 / 848 (33,0 %)
Proportion d’ensembles {1}	104 / 848 (12,3 %)
Proportion d’ensembles {0,1}	464 / 848 (54,7 %)
Proportion d’ensembles vides (∅)	0 / 848 (0,0 %)
Taille moyenne des ensembles Γ(x)	1,55

Analyse de sensibilité selon α (Mondrian)

Tableau . Analyse de sensibilité de la prédiction conforme Mondrian selon le niveau d’erreur α

α	Niveau de confiance (1–α)	Couverture globale	Taille moyenne Γ(x)	% ensembles {0}	% ensembles {1}	% ensembles {0,1}
0,05	95 %	0,958	1,79	15,4 % (131)	5,4 % (46)	79,1 % (671)
0,10	90 %	0,902	1,55	33,0 % (280)	12,3 % (104)	54,7 % (464)
0,20	80 %	0,777	1,25	50,5 % (428)	24,4 % (207)	25,1 % (213)
0,25	75 %	0,724	1,10	61,1 % (518)	29,2 % (248)	9,7 % (82)

Discussions et conclusion

Synthèse des principaux résultats

Ce travail avait pour objectif de comparer plusieurs stratégies de modélisation du risque de maladie coronarienne à 10 ans à partir de la base Framingham, en articulant trois dimensions méthodologiques majeures : gestion des données manquantes, traitement du déséquilibre de classes et quantification de l'incertitude. Dans cette cohorte de 4 239 sujets, la prévalence des événements coronariens est d'environ 15 %, configuration typique des études de prévention cardiovasculaire où les cas restent minoritaires. Sur le plan du prétraitement, la comparaison structurée de trois approches d'imputation (médiane/mode, KNN, MICE) montre que l'imputation simple, bien que facile à implémenter, tend à lisser les distributions, tandis que le KNN réduit de manière excessive la variance de plusieurs variables continues. À l'inverse, l'imputation multivariée par équations en chaîne (MICE) préserve mieux la distribution, la variance et les corrélations entre prédicteurs, en accord avec le rôle désormais central de cette méthode dans la gestion des données manquantes en épidémiologie et en modélisation prédictive (34). En l'absence de rééquilibrage, les modèles entraînés sur les jeux A1/A2 présentent une exactitude globale (accuracy) et une AUC-ROC correctes, mais un rappel extrêmement faible pour la classe événement, ce qui revient, en pratique, à manquer la majorité des patients qui feront un événement. L'introduction de SMOTE sur le jeu d'apprentissage (B1/B2) augmente systématiquement le rappel et le F1 de la classe 1, au prix d'une baisse modérée de l'accuracy, ce qui est conforme aux effets attendus du sur-échantillonnage dans les contextes de données fortement déséquilibrées (35). L'analyse détaillée montre qu'un modèle de régression logistique entraîné sur données MICE standardisées, sans SMOTE (LogReg_A2), mais associé à un **seuil de décision optimisé à 0,15**, offre un compromis particulièrement intéressant : amélioration nette du rappel (≈ 60 %) et du F1 de la classe 1, au prix d'une baisse acceptable de l'accuracy, tout en conservant une discrimination modérée mais réelle (AUC-ROC ≈ 0,70 ; AUC-PR ≈ 0,29) et un Brier score compatible avec une calibration globalement satisfaisante. Les analyses par Bootstrap confirment la stabilité de ces métriques, avec des intervalles de confiance relativement resserrés, et la prédiction conforme Mondrian fournit des ensembles prédictifs respectant globalement le niveau de couverture cible (~90 %), tout en signalant explicitement les situations d'incertitude élevée.

Comparaison avec la littérature et place de la régression logistique

Historiquement, les scores de risque dérivés de la Framingham Heart Study reposent sur des modèles de régression multivariée (logistique ou de Cox) intégrant les facteurs de risque classiques (âge, sexe, pression artérielle, profil lipidique, tabac, diabète). Ces fonctions de risque constituent toujours une référence dans les recommandations de prévention cardiovasculaire, même si plusieurs travaux ont souligné leurs limites de transportabilité et la nécessité de recalibrages locaux (36). Le profil des coefficients du modèle LogReg_A2 (effet dominant de l'âge, sur-risque masculin, rôle de la pression systolique, du nombre de cigarettes, du cholestérol et de la glycémie) est cohérent avec les déterminants décrits dans les études fondatrices de Framingham et dans les ouvrages de référence sur les modèles de prédiction clinique. Dans la littérature récente, plusieurs analyses comparatives suggèrent que, sur des données cardiovasculaires structurées et moyennement dimensionnelles, une régression logistique bien spécifiée, correctement imputée et validée, peut atteindre des performances proches de celles de méthodes plus complexes (forêts aléatoires, gradient boosting, XGBoost, SVM), tout en offrant une meilleure transparence et un contrôle plus fin de la calibration (37).

Nos résultats s'inscrivent dans cette ligne : les modèles d'arbres et d'ensembles montrent parfois des performances élevées en validation croisée sur le train, en particulier lorsqu'ils sont combinés à SMOTE, mais leur avantage en termes d'AUC sur le test réel reste limité, tandis que le risque de sur-apprentissage augmente dans le contexte de sur-échantillonnage. À l'inverse, la régression logistique appliquée au jeu A2, associée à un ajustement explicite du seuil de décision, apparaît comme un **compromis robuste** : discrimination et rappel cliniquement pertinents, structure paramétrique interprétable, et compatibilité avec les approches classiques de score de risque. Il ne s'agit pas de conclure que la régression logistique serait "universellement supérieure" aux méthodes d'apprentissage automatique, mais plutôt de montrer que, dans un cadre rigoureusement préparé, elle demeure un candidat très compétitif et scientifiquement défendable (38).

Forces méthodologiques et apport de la quantification de l'incertitude

Plusieurs éléments méthodologiques renforcent la robustesse de ce travail. D'abord, la gestion des valeurs manquantes repose sur une comparaison formalisée de trois méthodes d'imputation, incluant diagnostics graphiques, tests (Kolmogorov–Smirnov, Levene), évaluation de la variance par Bootstrap et vérification de l'impact sur la capacité prédictive. Cette démarche est en phase avec les recommandations de van Buuren, Steyerberg et d'autres auteurs qui insistent sur l'importance de considérer la stratégie d'imputation comme une

composante à part entière du développement des modèles de prédiction. Ensuite, la séparation stricte apprentissage/test avant toute opération de prétraitement, l'usage d'une validation croisée stratifiée et la quantification de l'incertitude par Bootstrap s'inscrivent dans l'esprit des lignes directrices TRIPOD et TRIPOD+AI, qui visent à améliorer la transparence, la reproductibilité et la qualité méthodologique des études de modèles prédictifs, qu'elles reposent sur des régressions classiques ou sur des méthodes d'apprentissage automatique(39). Ces lignes directrices recommandent notamment (i) de documenter le flux de données (split, imputation, rééchantillonnage), (ii) de distinguer clairement validation interne et externe, et (iii) de rapporter des mesures d'incertitude sur les performances, ce qui est fait ici via le bootstrap.

Enfin, l'intégration d'une prédiction conforme Mondrian autour du modèle final constitue un apport original. Cette approche fournit des ensembles prédictifs assortis de garanties de couverture finie, conditionnellement aux classes, sans hypothèse forte sur la forme du modèle.

Dans notre application, un niveau de confiance cible de 90 % conduit à une couverture empirique proche de l'objectif, avec un compromis intéressant entre singletons (prédictions "sûres") et ensembles {0,1} (cas explicitement signalés comme incertains). Cette capacité à signaler les situations où le modèle "hésite" est particulièrement pertinente dans une optique d'aide à la décision clinique, où l'acceptation d'une sortie "incertaine" peut être plus sûr que l'illusion d'une classification binaire certaine (40).

Limites et incertitudes persistantes

Ce projet présente toutefois plusieurs limites importantes. Sur le plan des données, la base Framingham utilisée correspond à une cohorte historique, majoritairement issue d'une population nord-américaine blanche, ce qui limite d'emblée la généralisable à d'autres contextes géographiques, ethniques ou de prise en charge. De nombreuses études ont montré que les scores dérivés de Framingham peuvent être mal calibrés lorsqu'ils sont appliqués à d'autres populations, nécessitant un recalibrage local voire un redéveloppement du modèle.

Dans ce travail, seule une **validation interne** (split apprentissage/test, validation croisée, bootstrap) a été réalisée ; aucune validation externe n'a pu être conduite, ce qui empêche toute conclusion directe sur la transportabilité et la calibration hors de la cohorte d'origine.

Par ailleurs, le choix de SMOTE comme principal outil de gestion du déséquilibre de classes ne couvre qu'une partie des stratégies disponibles. D'autres approches – variantes de SMOTE combinées à des méthodes de nettoyage (SMOTE-ENN, SMOTETomek), sous-échantillonnage de la classe majoritaire, pondération de la fonction de coût – pourraient produire des compromis différents entre rappel, précision, AUC et calibration (35).

De plus, nos résultats montrent que certains modèles d'arbres ou d'ensembles sur-apprennent facilement lorsqu'ils sont entraînés sur des données sur-échantillonnées, ce qui invite à la prudence dans l'interprétation des performances élevées observées en validation croisée sur le train.

Enfin, la quantification de l'incertitude proposée ici, bien que plus riche que dans de nombreux travaux (bootstrap des métriques, prédiction conforme des sorties individuelles), demeure partielle. Elle ne capture pas l'incertitude liée au choix de l'algorithme, aux décisions de prétraitement (par exemple, choix du schéma d'imputation ou de la stratégie de rééquilibrage), ni les effets potentiels d'un décalage de distribution entre la population de Framingham et la population cible réelle. La littérature récente sur les modèles "uncertainty-aware" insiste sur la nécessité de combiner plusieurs sources d'incertitude (données, modèle, scénario d'application) et de les communiquer de manière intelligible aux cliniciens (41,42).

Implications cliniques et perspectives de recherche

Sur le plan clinique, les résultats suggèrent qu'un modèle de type Framingham ré-estimé et modernisé – basé sur une régression logistique multivariée, entraînée après imputation MICE, standardisation et prise en compte explicite du déséquilibre via le choix du seuil – pourrait constituer un score de risque opérationnel pour la prévention cardiovasculaire. Ce modèle reste conceptuellement proche des outils actuellement utilisés (Framingham, SCORE2), tout en intégrant des standards méthodologiques plus récents en matière de gestion des données manquantes, de validation interne et de reporting. Le compromis observé (légère baisse d'accuracy globale mais gain substantiel de rappel) est cohérent avec une logique clinique où l'on préfère détecter davantage de patients à haut risque, quitte à accepter davantage de faux positifs qui pourront être re-classés lors d'une évaluation clinique approfondie. La prédiction conforme offre, en complément, un moyen explicite de distinguer les cas pour lesquels le modèle est relativement confiant (ensembles {0} ou {1}) de ceux pour lesquels l'incertitude est élevée (ensembles {0,1}). À terme, cette information pourrait être intégrée à une interface

clinique de type “feu tricolore” (vert = faible risque, rouge = haut risque, orange = incertitude élevée), conformément aux appels récents en faveur de modèles prédictifs capables de rendre compte de leur propre incertitude plutôt que de fournir des scores “muets” (43). Plusieurs prolongements de recherche apparaissent naturels. D’un point de vue méthodologique, il serait pertinent de comparer systématiquement différentes stratégies de rééquilibrage (SMOTE-ENN, SMOTETomek, pondération des classes, méthodes focales) dans le même cadre, en évaluant conjointement discrimination, calibration et impact clinique potentiel (35,37). D’un point de vue applicatif, une validation externe sur des cohortes plus diverses, idéalement en suivant les recommandations TRIPOD/TRIPOD+AI pour le développement, la validation et le reporting, constitue une étape indispensable avant toute implémentation opérationnelle (44). Enfin, l’intégration plus large de méthodes d’incertitude (approches bayésiennes, ensembles de modèles, combinaisons avec la conformal prediction) pourrait être explorée afin d’aligner au mieux les sorties des modèles avec les besoins de la décision partagée entre clinicien et patient.

Conclusion

Ce travail avait pour objectif de développer et comparer plusieurs modèles de prédiction du risque de maladie coronarienne à 10 ans à partir de la base Framingham, en accordant une attention particulière à trois dimensions souvent critiques mais parfois insuffisamment explicitées dans la pratique : la gestion des valeurs manquantes, le déséquilibre de classes et la quantification de l’incertitude. Sur le plan du prétraitement, l’imputation multivariée par MICE s’est imposée comme la stratégie la plus satisfaisante pour préserver les distributions, la variance et les corrélations entre prédicteurs, justifiant son adoption comme méthode de référence pour la constitution des jeux d’apprentissage et de test (45). L’analyse des modèles a montré que, sans rééquilibrage des classes, les performances globales peuvent paraître correctes (accuracy, AUC-ROC), tout en masquant une sensibilité très insuffisante pour la classe événement, ce qui limite fortement l’intérêt clinique. L’utilisation de SMOTE sur le jeu d’apprentissage améliore nettement le rappel et le F1 de la classe 1 pour l’ensemble des algorithmes, en illustrant le compromis classique entre détection des cas et exactitude globale dans des données fortement déséquilibrées.

Dans ce cadre, une régression logistique multivariée entraînée sur données imputées et standardisées, sans SMOTE mais avec un seuil de décision optimisé (LogReg_A2, seuil 0,15), apparaît comme un modèle de compromis : elle offre une sensibilité accrue pour les événements, une discrimination comparable aux modèles d'arbres et d'ensembles, et conserve une structure interprétable, en continuité avec les approches classiques de score de risque en épidémiologie clinique. La quantification de l'incertitude par bootstrap et prédiction conforme montre que ces performances sont relativement robustes aux fluctuations d'échantillonnage et permet d'associer à chaque prédiction un niveau de confiance explicite, élément important en vue d'une possible intégration dans un système d'aide à la décision.

Ce travail doit toutefois être envisagé comme une étape exploratoire : il repose sur une seule cohorte historique, n'intègre qu'une validation interne et n'épuise ni l'espace des stratégies de rééquilibrage, ni celui des méthodes de quantification de l'incertitude. Une validation externe sur des populations plus diverses, un recalibrage éventuel et une comparaison élargie des approches de gestion du déséquilibre seront indispensables avant toute utilisation clinique. Néanmoins, il illustre qu'un pipeline rigoureux combinant analyse exploratoire, imputation multivariée, gestion structurée du déséquilibre, validation conforme aux recommandations TRIPOD/TRIPOD+AI et outils modernes d'incertitude – permet d'articuler de manière cohérente méthodes d'apprentissage automatique et exigences de transparence, dans la perspective d'une prédiction du risque cardiovasculaire réellement utile à la décision médicale.

Références biographiques

1. Cardiovascular diseases (CVDs) [Internet]. [cité 16 janv 2026]. Disponible sur: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

2. Charge mondiale de morbidité liée aux cardiopathies ischémiques de 2022 à 2050 : projections de l'incidence, de la prévalence, de la mortalité et des années de vie corrigées de l'incapacité | European Heart Journal - Quality of Care and Clinical Outcomes | Oxford Academic [Internet]. [cité 16 janv 2026]. Disponible sur: https://academic.oup.com/ehjqcco/article/11/4/355/7699087?utm_source=chatgpt.com

3. Maladies cardiovasculaires (risque à 10 ans) | Étude de Framingham sur le cœur [Internet]. [cité 16 janv 2026]. Disponible sur: https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk/?utm_source=chatgpt.com

4. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 7 janv 2015;350:g7594.

5. Dehghan A, Jahangiry L, Khezri R, Jafari A, Pezeshki B, Rezaei F, et al. Framingham risk scores for determination the 10-year risk of cardiovascular disease in participants with and without the metabolic syndrome: results of the Fasa Persian cohort study. BMC Endocr Disord. 24 juin 2024;24:95.

6. Papangelou C, Kyriakidis K, Natsiavas P, Chouvarda I, Malousi A. Reliable machine learning models in genomic medicine using conformal prediction. Front Bioinforma. 24 févr 2025;5:1507448.

7. Jaber E, Blot V, Brunel NJB, Chabridon V, Remy E, Iooss B, et al. CONFORMAL APPROACH TO GAUSSIAN PROCESS SURROGATE EVALUATION WITH MARGINAL COVERAGE GUARANTEES. 2025;6(3).

8. Royston P, White IR. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. J Stat Softw. 12 déc 2011;45:1-20.

9. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 6 janv 2015;162(1):W1-73.

10. Papangelou C, Kyriakidis K, Natsiavas P, Chouvarda I, Malousi A. Reliable machine learning models in genomic medicine using conformal prediction. Front Bioinforma. 24 févr 2025;5:1507448.

11. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health): 9783030163983: Medicine & Health Science Books @ Amazon.com [Internet]. [cité 16 janv 2026]. Disponible sur: https://www.amazon.com/Clinical-Prediction-Models-Development-Validation/dp/3030163989?utm_source=chatgpt.com

12. Cardiovascular Risk Factors. Insights From Framingham Heart Study. Rev Esp Cardiol Engl Ed. 1 mars 2008;61(3):299-310.

13. Steyerberg EW. Validation of Prediction Models. In: Steyerberg EW, éditeur. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating [Internet]. Cham: Springer International Publishing; 2019 [cité 16 janv 2026]. p. 329-44. Disponible sur: https://doi.org/10.1007/978-3-030-16399-0_17

14. Zhang Z. Prediction of heart disease based on logistic regression. Theor Nat Sci. 31 août 2024;51:1-7.

15. Steyerberg EW. Statistical Models for Prediction. In: Steyerberg EW, éditeur. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating [Internet]. Cham: Springer International Publishing; 2019 [cité 16 janv 2026]. p. 59-93. Disponible sur: https://doi.org/10.1007/978-3-030-16399-0_4

16. https://stefvanbuuren.name/fimd/index.html [Internet]. [cité 16 janv 2026]. Disponible sur: https://stefvanbuuren.name/fimd/index.html

17. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 12 déc 2011;45:1-67.

18. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 20 févr 2011;30(4):377-99.

19. Little RJA, Rubin DB. Statistical Analysis with Missing Data. John Wiley & Sons; 2019. 462 p.

20. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ. 3 sept 2024;386:e078276.

21. He H, Garcia EA. Learning from Imbalanced Data. Knowl Data Eng IEEE Trans On. 1 oct 2009;21:1263-84.

22. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 1 juin 2002;16:321-57.

23. Steyerberg E. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Vol. 19. 2009.

24. Breiman L. Random Forests. Mach Learn. 1 oct 2001;45(1):5-32.

25. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2016 [cité 16 janv 2026]. p. 785-94. (KDD '16). Disponible sur: https://dl.acm.org/doi/10.1145/2939672.2939785

26. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 4 mars 2015;10(3):e0118432.

27. Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Mon Weather Rev. 1 janv 1950;78(1):1-3.

28. Smith GCS, Seaman SR, Wood AM, Royston P, White IR. Correcting for Optimistic Prediction in Small Data Sets. Am J Epidemiol. 1 août 2014;180(3):318-24.

29. Efron B. Introduction à Bootstrap. 456 p.

30. Farooq V, Brugaletta S, Vranckx P, Serruys PW. A guide to interpreting and assessing the performance of prediction models [Internet]. [cité 16 janv 2026]. Disponible sur: https://eurointervention.pcronline.com/article/a-guide-to-interpreting-and-assessing-the-performance-of-prediction-models

31. Assel M, Sjoberg DD, Vickers AJ. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagn Progn Res. 2 déc 2017;1:19.

32. Shafer G, Vovk V. A tutorial on conformal prediction [Internet]. arXiv; 2007 [cité 16 janv 2026]. Disponible sur: http://arxiv.org/abs/0706.3188

33. Ding T, Angelopoulos AN, Bates S, Jordan MI, Tibshirani RJ. Class-Conditional Conformal Prediction with Many Classes.

34. Buuren S van, Groothuis-Oudshoorn CGM. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw [Internet]. 2011 [cité 16 janv 2026];45(3). Disponible sur: https://research.utwente.nl/en/publications/mice-multivariate-imputation-by-chained-equations-in-r/

35. He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng. 1 sept 2009;21(9):1263-84.

36. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating | Springer Nature Link [Internet]. [cité 16 janv 2026]. Disponible sur: https://link.springer.com/book/10.1007/978-0-387-77244-8?referer=www.springeronline.com&utm_source=chatgpt.com

37. Kwakye K, Dadzie E. Machine Learning-Based Classification Algorithms for the Prediction of Coronary Heart Diseases [Internet]. arXiv; 2021 [cité 16 janv 2026]. Disponible sur: http://arxiv.org/abs/2112.01503

38. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc JAMIA. 16 août 2022;29(9):1525-34.

39. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 7 janv 2015;350:g7594.

40. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data | Journal of Big Data | Springer Nature Link [Internet]. [cité 16 janv 2026]. Disponible sur: https://link.springer.com/article/10.1186/s40537-023-00857-7?utm_source=chatgpt.com

41. Hoblos R, Dridi N, Zerhouni N, Masry ZA. An Uncertainty Quantification Method Based on Evidence theory and Conformal Prediction.

42. Gade M, Nguyen KM, Gedde S, Fernandez-Quilez A. Impact of uncertainty quantification through conformal prediction on volume assessment from deep learning-based MRI prostate segmentation. Insights Imaging. 29 nov 2024;15(1):286.

43. 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. 16 avr 2024;385:e078378.

44. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 7 janv 2015;350:g7594.

45. Buuren S van, Groothuis-Oudshoorn CGM. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw [Internet]. 2011 [cité 16 janv 2026];45(3). Disponible sur: https://research.utwente.nl/en/publications/mice-multivariate-imputation-by-chained-equations-in-r/

Annexes