CHAPITRE 4 – SCORE DE RISQUE POLYGÉNIQUE

4.1 Génétique à l'ère post-GWAS

Avec l'accroissement rapide des données génétiques des dernières années, une quantité substantielle d'informations est maintenant disponible et accessible par les chercheurs en génomique. Les études GWAS ont révolutionné la génétique médicale en mettant en évidence des éléments génomiques reliés à certaines maladies. Comme les études GWAS sont puissantes dans la détection des variations génomiques impliquées dans les traits quantitatifs chez plusieurs individus elles peuvent donc continuer de servir de point d'ancrage dans une ère dite post-GWAS (analyse de données brutes) [30]. Toutefois, l'héritabilité expliquée dans les études GWAS demeure relativement faible dans le cas des traits complexes partiellement en raison des diverses interactions gènes-environnement. En prenant en considération tous les loci susceptibles d'avoir un impact sur le phénotype clinique, il est possible d'établir un outil statistique faisant la somme pondérée de l'effet de chacun de ces locus [149]. Cette méthode est nommée score de risque polygénique et présente l'avantage de pouvoir être utilisée comme outil de prédiction pour des pathologies dans l'optique de définir la prédisposition qu'a un individu face à une maladie et d'en prévenir le développement pouvant affecter de manière considérable la vie des patients (maladies dégénératives, cancers héréditaires ou encore des maladies mentales envahissantes). Plus le nombre de loci inclus dans le PRS est important, plus la signification clinique de ce dernier augmente. La forme la plus communément utilisée pour l'établissement d'un PRS consiste à effectuer la somme des m allèles de susceptibilité multipliée par la taille de l'effet β (ln(OR)) [150].

$$PRS_i = \sum_{j=1}^{m} x_{ij}\beta_j$$

Plusieurs considérations sont à prendre lors de l'utilisation et de l'établissement d'un PRS. En premier lieu, l'emploi de grand nombre de SNP dans la construction du modèle peut avoir un effet plus négatif que positif sur la prédictibilité en augmentant l'ampleur de la taille de l'effet induisant ainsi un « bruit de fond » dans le modèle. [150]. Un choix judicieux des SNP à être inclus dans le score s'impose en se basant sur l'effet plus important et significatif (p-value). En second lieu, le déséquilibre de liaison (**DL**) entre des SNP voisins résultant en une surreprésentation de certains loci dans le modèle ayant pour effet de réduire son pouvoir discriminatoire. Pour faire face à ces problèmes, un SNP aléatoire d'une région en DL est retiré du modèle ou le seuil de valeur p est augmenté pour prendre en compte moins de SNP en DL [150, 151]. L'engouement ressenti depuis quelques années pour le développement de PRS repose sans doute sur sa capacité à prédire le développement des traits complexes à prévalence faible mieux que l'histoire familiale qui réussit seulement à expliquer environ 4% de l'héritabilité des traits quantitatifs [152]. Un autre avantage des PRS est celui de la prédiction du risque de la maladie à long terme (*lifetime risk trajectory*). En effet, en utilisant une courbe de Kaplan-Meier à partir des données du PRS, il est possible de déterminer le risque cumulatif de développer la maladie en fonction de l'âge des individus (Figure 4, où la courbe grise représente le risque moyen, la courbe rouge le risque plus élevé et la courbe bleu le risque le moins élevé).

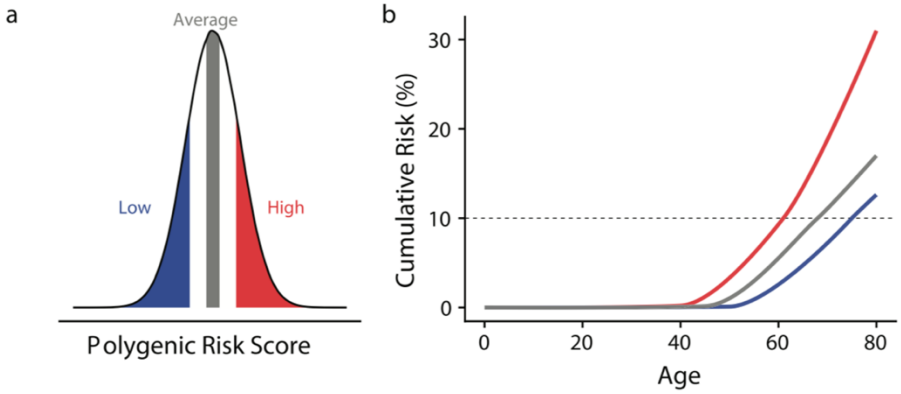

Figure 5 Courbe de Kaplan-Meier du risque cumulatif (%) en fonction de l'âge [tiré de 150]

Avec les mouvements migratoires de plus en fréquents et facilités depuis les dernières décennies, les populations de certains pays présentent une composante multiculturelle de plus en plus prononcée. Cependant, en 2019, il est estimé qu'environ 79% de toutes les études GWAS sont basées sur des individus d'origine européenne en dépit du fait qu'il ne représente que 16% de la population mondiale réduisant ainsi le nombre de PRS appliqués à d'autres populations [153]. Les PRS développés de 2008 à 2017 surreprésentent les populations caucasiennes (459 études) tandis que seulement 140 études concernent les populations asiatiques, 15 études sur les populations africaines, 9 sur les populations latino-américaines et 4 sur celles du Moyen-Orient [154]. Dans l'optique d'adapter les méthodes de recherche pour les rendre plus universelles et accessibles à toutes les populations, l'établissement d'un PRS multiethnique s'avère donc essentiel. Pour augmenter le pouvoir de prédiction du score de risque polygénique, des chercheurs ont utilisé des données GWAS d'origine européenne combinées avec des données GWAS issues de populations latino-américaines pour augmenter la prédictibilité de plus de 70% pour le diabète de type II dans la population sud-américaine [155]. Une autre équipe de recherche américaine s'est penchée sur l'étude de la dermite atopique dans deux cohortes et comportant plus de 90 000 individus d'origine afro-américaine établis aux États-Unis. Ils ont développé un score de risque polygénique basé sur 27 SNP pangénomiques identifiés dans la population caucasienne et ont conclu que le PRS développé était hautement prédictif de la dermite atopique dans la population caucasienne mais demeurait très peu prédictif chez les afro-américains suggérant ainsi une architecture génétique différente pour la dermite atopique dans les deux populations [156]. Aujourd'hui, de nombreux PRS ont été établi pour diverses maladies et plusieurs d'entre eux ont effectivement permis d'effectuer des mesures préventives pour contrer le développement de maladies invasives soulignant donc leur importance dans une ère de médecine dite « personnalisée » [157-160]. L'exemple particulier d'un PRS établi pour le cancer du sein au Royaume-Uni chez 67 000 femmes d'origine européenne a permis

d'identifier les femmes à risque de développer un cancer du sein dix ans avant la moyenne des femmes et donc dix ans avant le début des tests de dépistage standards [161]. Toutefois, des efforts sont toujours requis pour mieux adapter ces outils de prédiction et les rendre plus universels pour ainsi améliorer la prévention.

4.2 Fonction d'efficacité du récepteur

Les courbes de fonction d'efficacité du récepteur ou plus communément appelées, en français, courbes **ROC** (*Receiver Operating Characteristic*) sont des outils graphiques mettant en relation la spécificité et la sensitivité d'un modèle phénotypique dichotomique (patients atteints ou non atteints) et continu dans l'optique de mesurer la performance. Une courbe ROC typique est séparée en deux portions, par la ligne d'égalité des chances, qui comprennent chacune une aire sous la courbe (**AUC**) de 0,5. Si la courbe se trouve sur la ligne d'égalité des chances, l'AUC est, par conséquent, de 0,5 ce qui signifie que le modèle prédictif développé discrimine les individus atteints des non atteints dans 50% des cas (Figure 6) [162]. La mesure d'AUC résume la position globale de la courbe et plus sa mesure est élevée (le plus près de 1), plus le modèle diagnostic fait la distinction entre patients et témoins [163]. Statistiquement, il est plus avantageux de retenir la mesure d'AUC comme seuil de prédiction que des points quelconques car elle représente la sensitivité moyenne pour toutes les valeurs de spécificité possibles [163, 164]. Plusieurs avantages sont reliés à l'emploi de courbes ROC d'un point de vue statistique et médical. Les courbes ROC sont indépendantes de la prévalence de la pathologie étudiée car elles se basent sur les mesures de sensitivité et de spécificité de l'outil de prédiction. Toutefois, si la prévalence est relativement faible, le taux de faux positifs (1 – spécificité, correspondant à l'axe des X) devra, par ailleurs, demeurer bas sinon tous les patients classés comme atteints seront en réalité non atteints et vice-versa [165].

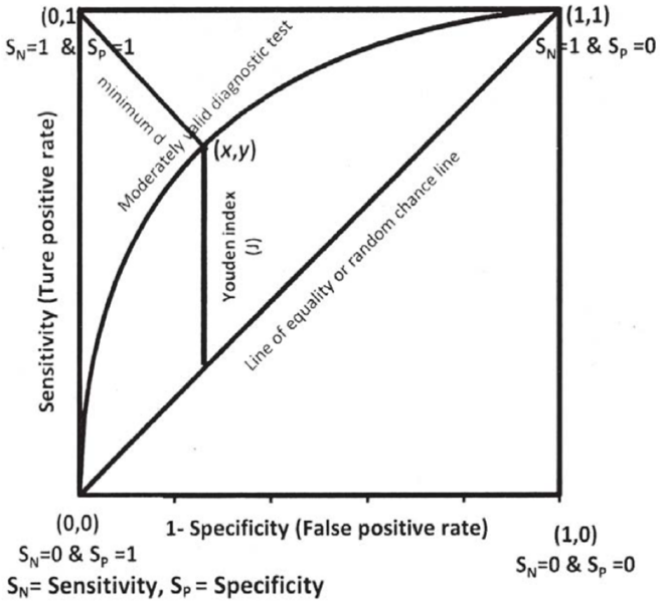

Figure 6 Structure d'une courbe ROC [tiré de 162]

La sensitivité, la spécificité ainsi que l'exactitude sont calculées à l'aide des équations suivantes (note : **NT** représente le nombre total d'individus; **NTA** le nombre de patients réellement atteints (donc diagnostiqués) et **NTN** le nombre de patients réellement non atteints) :

$$Sensibilité = \frac{proportion\ de\ vrais\ positifs\ (VP)}{proportion\ de\ vrais\ positifs\ (VP) + proportion\ de\ faux\ négatifs\ (FN)}$$

$$Spécificité = \frac{proportion\ de\ vrais\ négatifs\ (VN)}{proportion\ de\ faux\ positifs\ (FP) + proportion\ de\ vrais\ négatifs\ (VN)}$$

$$Exactitude = \frac{VP}{NTA}\ X\ \frac{NTA}{NT} + \frac{VN}{NTN} X\ \frac{NTN}{NT}$$

Plusieurs outils statistiques ont été développés en parallèle à la courbe ROC pour tenir compte et essayer d'expliquer la variance dans les modèles logistiques. Un peu à la manière du R^2 dans une régression linéaire renseigne sur l'adaptabilité du modèle en fonction des variables choisies ou, autrement, à quel point le modèle décrit la tendance linéaire des données sur une échelle de 0 à 1 (0 correspondant à un modèle n'expliquant pas la tendance linéaire et 1 un modèle qui l'explique à 100%). L'équivalent du R^2 de la régression linéaire dans un modèle logistique est le pseudo-R^2. Néanmoins, il de mise de rester prudent quant à son utilisation car il ne correspond pas exactement au R^2 des régressions linéaires mais bien à une probabilité, ce pourquoi il est nommé « pseudo » R^2 [166]. De nombreux statisticiens ont établi des mesures de pseudo-R^2 dont McFadden, Cox et Snell et Cragg et Uhler (repris par Nagelkerke d'où l'ambiguïté sur le nom officiel de cette statistique). L'une de ces statistiques la plus employée lors d'études épidémiologiques est celle de Nagelkerke. Il permet de calculer le pourcentage de variance expliquée par le modèle. Le pseudo-R^2 de Nagelkerke correspond à celui de Cox-Snell corrigé par le maximum de la mesure du pseudo-R^2 de Cox-Snell [166]. Il se calcule à partir de l'équation suivante (note : L correspond au logarithme de la vraisemblance, M_0 au modèle sans variables, M_1 au modèle avec variables et N au nombre total d'observations) :

$$R^2_{Nagelkerke} = \frac{R^2_{Snell-Cox}}{R^2_{Snell-Cox}\ maximal} = \frac{1 - \left[\frac{L(M_0)}{L(M_1)}\right]^{\frac{2}{N}}}{1 - L(M_0)^{\frac{2}{N}}}$$

Il est de mise d'être prudent lorsque le R^2 de Nagelkerke est employé car il a tendance à maximiser les valeurs de R^2 comparativement aux autres mesures. Il a aussi tendance à prendre des valeurs d'ajustement plus grandes qu'en réalité [166]. Plus récemment, d'autres méthodes statistiques ont été développées pour évaluer l'efficacité du modèle de prédiction. Le *Net Reclassification Index* (**NRI**) compare un ancien modèle de prédiction avec un nouveau modèle prédictif. Lorsqu'un nouveau marqueur est ajouté au modèle le NRI mesure les patients qui seront reclassés comme non atteints s'ils étaient malades ou les témoins qui seront considérés comme malades si l'ajout de marqueur influence la maladie [167].