L'Intelligence Artificielle au service de la médecine de précision en transplantation

Marc Labriffe

Thèse de doctorat

Université de Limoges

Université de Limoges

ED 652 - Biologie, Chimie, Santé (BCS)

Unité Inserm 1248 Pharmacologie & Transplantation

Thèse pour obtenir le grade de

Docteur de l'Université de Limoges

Pharmacologie, infectiologie et sciences du médicament

Présentée et soutenue par

Marc Labriffe

Le 23 octobre 2023

L'intelligence artificielle au service de la médecine de précision en transplantation

Thèse dirigée par Professeur Pierre Marquet

JURY :

Présidente du jury
Mme Magali Giral, PU-PH, Université de Nantes

Rapporteurs
M. Rodolphe Thiébaut, PU-PH, Université de Bordeaux
Mme Claire Tinel, PU-PH, Université de Dijon

Examinateurs
M. Pierre Marquet, PU-PH, Université de Limoges
M. Jean-Baptiste Woillard, PU-PH, Université de Limoges
Mme Aurélie Prémaud, MCU, Université de Limoges

Remerciements

Tout d’abord, je remercie vivement le Professeur Pierre Marquet d’avoir encadré cette thèse. Vous m’avez fait confiance et cela me permet de participer à de nombreux projets de recherche dans le service. Je vous suis également reconnaissant pour le temps que vous m’avez accordé à relire et corriger tous mes travaux.

Je tiens ensuite à remercier le Professeur Rodolphe Thiébaut et la Professeure Claire Tinel. Vous me faites l’honneur d’être les rapporteurs de ce travail et de partager votre expertise en biostatistiques et néphrologie. Veuillez trouver ici l’expression de ma sincère gratitude.

Je remercie ensuite la Professeure Magali Giral. Veuillez recevoir ma profonde reconnaissance pour avoir accepté d’examiner ce travail de thèse. Merci aussi au Docteur Aurélie Prémaud d’avoir accepté de participer à ce jury.

Impossible de ne pas remercier le Professeur Jean-Baptiste Woillard, pour m’avoir encouragé à coder, et à utiliser du Machine Learning dans certains de mes projets. Merci pour ton enthousiasme et ton humour !

Je n’oublie pas non plus Antoine Humeau, pour l’aide qu’il m’a apportée dans le nettoyage et la préparation de certaines bases de données.

Je remercie chaleureusement Franck Saint-Marcoux et Caroline Monchaud, merci pour votre écoute, je profite quotidiennement de vos avis. Merci à Souleiman El Balkhi, Sylvain Couderc, et Nicolas Picard, votre aide dans le service m’est également très précieuse.

Droits d’auteurs

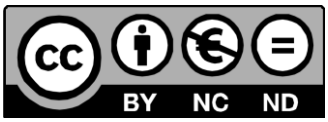


Table des matières

Abréviations

ABIS	Site internet d'Adaptation Bayésienne des ImmunoSuppresseurs https://abis.chu-limoges.fr
ABMR	Rejet médié par les anticorps
ah	Score de Banff de hyalinose artériolaire
AdGFS	Adjustable graft failure score
AMP	Acide mycophénolique
AUC	*Area under the curve*, aire sous la courbe (ASC)
BKV	Virus BK
C4d	Produit inactif, issu de la dégradation catalytique du complément C4
cg	Score de Banff pour les doubles contours sur la membrane basale glomérulaire
ci	Score de Banff de fibrose interstitielle
CMV	Cytomégalovirus
CNI	Inhibiteur de la calcineurine
CNN	Réseaux de neurones convolutifs
ct	Score de Banff d'atrophie tubulaire
DFGe	Débit de filtration glomérulaire estimé
DSA	Anticorps anti-HLA du donneur
EMA	Agence européenne des médicaments
g	Score de Banff pour la glomérulite
HLA	Antigènes des leucocytes humains
HPLC	Chromatographie en phase liquide à haute performance
i	Score de Banff d'inflammation dans l'interstitium
i-IFTA	Score de Banff d'inflammation dans la fibrose
IA	Intelligence artificielle
IFTA	Fibrose interstitielle et atrophie tubulaire
IHC	Immunohistochimie
MAP	Maximum a posteriori
ML	*Machine learning,* apprentissage automatique
MS	Syndrome métabolique
MMF	Mycophénolate mofétil
MVI	Invasion microvasculaire
mTOR	mammalian target of rapamycin

mTori	Inhibiteurs de la mammalian target of rapamycin
NDSA	Anticorps anti-HLA non spécifiques du donneur
PopPK	Pharmacocinétique de population
PR	Courbe precision en ordonnée (taux de vrais positifs ou valeur prédictive positive ou VPP) et recall en abscisse (sensibilité)
ptc	Score de Banff de capillarite, présence de cellules dans les capillaires péritubulaires
ptcml	Score de Banff pour la multilamellation de la membrane basale péritubulaire
PTDM	Diabète post transplantation
pvl	polyomavirus load level, score de Banff gradé de façon semi-quantitative en fonction du pourcentage de tubules cortico-médullaires sièges d'une réplication virale prouvée par immunohistochimie anti-SV40 et/ou par la présence d'inclusions virales dans au moins une cellules épithéliale tubulaire
Rexetris	Relations EXposition - Effet à long terme chez le Transplanté Rénal des médicaments ImmunoSuppresseurs
RMSE	Racine carrée de la moyenne des carrés des erreurs
sABMR	Suspicion de rejet médié par les anticorps
t	Score de Banff de tubulite
t-IFTA	Score de Banff de tubulite dans la fibrose
TCMR	Rejet médié par les cellules T
v	Score de Banff d'artérite intimale
XGBoost	eXtreme Gradient Boosting

Table des illustrations

Table des tableaux

Introduction

I. La médecine personnalisée en transplantation

La pharmacologie, c'est-à-dire l'étude des médicaments, de leur devenir dans l'organisme, de leurs effets et de leur emploi, a permis l'essor de la transplantation au cours des 60 dernières années [1,2]. En pratique clinique, son rôle est d'assurer au bon patient, l'apport du bon médicament, à la bonne posologie.

I.1. L'exposition au médicament

I.1.1. Suivi thérapeutique pharmacologique des traitements immunosuppresseurs en transplantation

Pour les immunosuppresseurs (IS), la mesure de l'exposition au médicament est essentielle afin de garantir son efficacité, tout en évitant des effets non désirés à long terme [3].

Pour la transplantation, cela concerne particulièrement :

- un antimétabolite, l'acide mycophénolique (AMP), métabolite actif du mycophénolate mofétil (MMF) qui inhibe de manière réversible et non compétitive une enzyme clé de la synthèse de novo des bases puriques, l'inosine monophosphate déshydrogénase (IMPDH) (Figure 1) ;

- les inhibiteurs de la calcineurine (CNI) : ciclosporine et tacrolimus, qui inhibent la translocation nucléaire du facteur de transcription NFAT[1] du gène de l'interleukine 2, et par conséquent la prolifération des lymphocytes T ;

- des inhibiteurs de mTOR : évérolimus et sirolimus, qui inhibent une protéine kinase qui favorise la croissance, la prolifération, la mobilité, ainsi que la survie cellulaire.

[1] NFAT : Nuclear factor of activated T-cells.

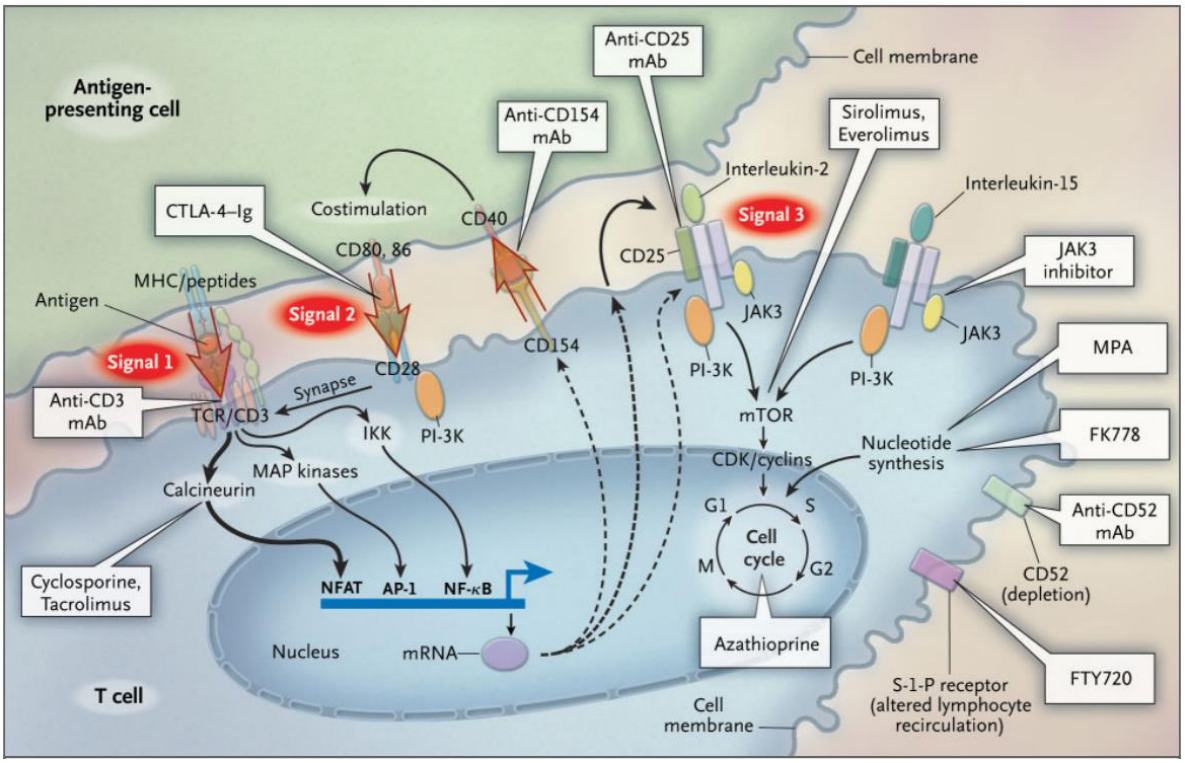


Figure 1 : Médicaments immunosuppresseurs en transplantation rénale et leur site d'action

Le suivi thérapeutique pharmacologique concerne particulièrement : l'acide mycophénolique (MPA, à droite), les inhibiteurs de la calcineurine (ciclosporine et tacrolimus, à gauche en bas), les inhibiteurs de mTOR (sirolimus et évérolimus, en haut à gauche).

Source : Halloran et al. [4]

Le suivi thérapeutique pharmacologique de ces médicaments immunosuppresseurs est en effet indiqué pour plusieurs raisons :

- zone thérapeutique étroite,
- variabilité pharmacocinétique interindividuelle importante,
- difficulté à mesurer l'effet clinique à court terme,
- survenue d'effets indésirables graves, parfois irréversibles et décelés après plusieurs années d'exposition,
- et bien sûr l'existence de méthodes de dosage.

L'AMP est un exemple de ces médicaments antirejets à fenêtre thérapeutique étroite. Pour l'AMP, l'objectif de l'aire sous la courbe des concentrations ($AUC_{0\text{-}12h}$) est de 30-60 mg.h/L chez les patients transplanté rénaux [5], avec une variabilité interindividuelle qui est importante [6]. Enfin, les méthodes de dosage utilisées sont multiples : chromatographie liquide [7], immunologiques [8], inhibition de l'IMPDH [9].

L'évérolimus en est un autre exemple. L'exposition à ce médicament peut être suivie par la mesure de la concentration résiduelle [10] ou par l'estimation de l'AUC à partir de plusieurs concentrations [11]. Il est aussi dosé par chromatographie liquide couplée à la spectrométrie de masse en tandem [12] qui est la méthode de référence pour ce médicament.

Parfois, la mesure objective de l'exposition n'est pas nécessaire. C'est le cas de médicaments dont la dose nécessaire est fixe chez tous les patients anticorps monoclonaux (ex. anti-CD25 – basiliximab –, bloqueurs de la costimulation du lymphocyte T – bélatacept –), plutôt déterminée par le poids du patient (ex. corticoïdes en traitement d'entretien), ou directement adaptée en fonction du résultat clinique (disparition des signes de rejet aigu, ex. corticoïdes en traitement du rejet aigu).

I.1.2. Estimation des aires sous la courbe des concentrations

L'expérience acquise dans l'utilisation du MMF a permis de montrer que la mesure de la concentration résiduelle ne suffisait pas pour estimer l'exposition à ce médicament [13], contrairement à l'estimation de l'AUC [14–17].

Pour l'évérolimus, à ce jour, l'apport de l'estimation de l'AUC a peu été étudié [18], bien que l'AUC soit théoriquement un meilleur marqueur de la pharmacodynamie du médicament que la concentration résiduelle. Kovarik et al. [11] ont mis en évidence la relation exposition-effet qui existe entre l'AUC de l'évérolimus et l'incidence de thrombopénie, hypertriglycéridémie et hypercholestérolémie. Les auteurs ont également remarqué que la variabilité interindividuelle de l'AUC était importante par rapport à la variabilité intra-individuelle, encourageant le suivi thérapeutique pharmacologique de ce médicament par l'AUC. Cette découverte a ensuite été confirmé dans d'autres études [19]. La raison principale de l'absence d'utilisation clinique de l'AUC de l'évérolimus est probablement la difficulté à construire des outils pour l'estimer de manière fiable.

Bien que plus informative, l'AUC est plus difficile à mesurer qu'une concentration résiduelle. En effet, le *gold standard*, la méthode des trapèzes, nécessite de réaliser plusieurs prélèvements à intervalles réguliers entre deux prises de médicament (Figure 2).

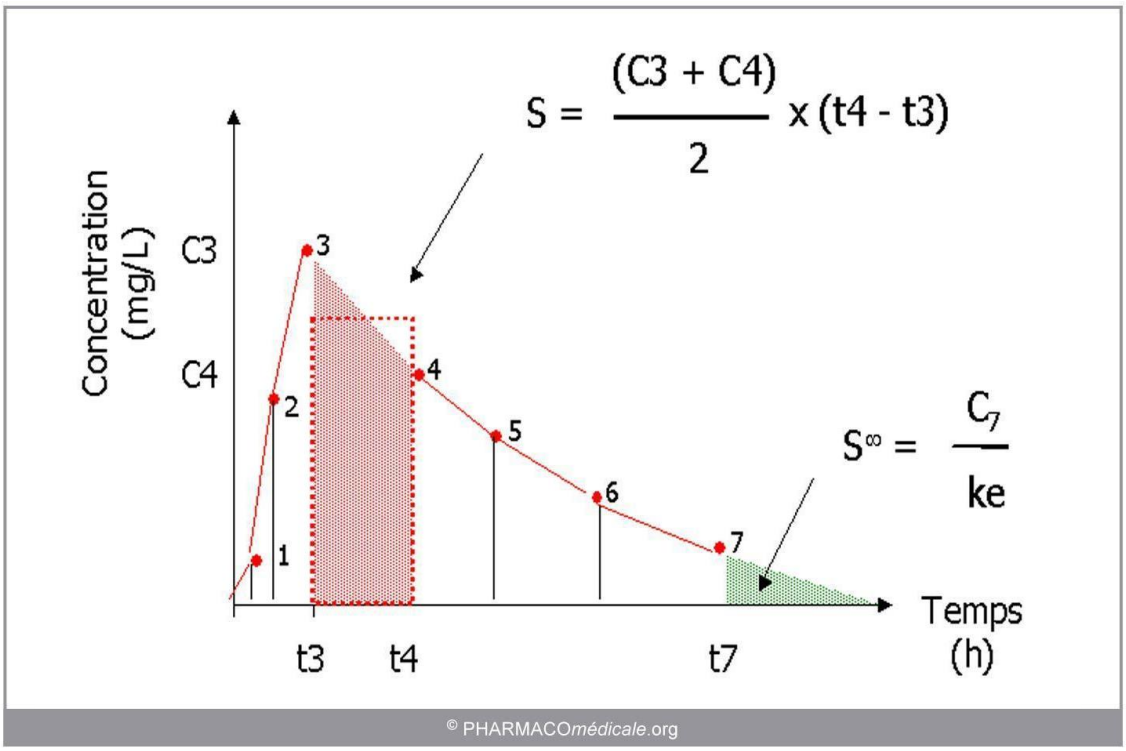


Figure 2 : Calcul de l'aire sous la courbe (AUC) des concentrations en fonction du temps par la méthode des trapèzes

L'aire d'un trapèze est généralement calculée en utilisant la concentration moyenne entre deux temps ($\frac{C3 + C4}{2}$), multipliée par l'intervalle entre ces deux temps ($t4 - t3$).

Source : pharmacomedicale.org [20]

En pratique, des stratégies d'échantillonnage limité ont été développées pour permettre d'estimer l'AUC de manière moins invasive, moins coûteuse et plus adaptée à une courte hospitalisation de jour. C'est le cas pour les modèles pharmacocinétiques de population (PopPK) utilisés avec des estimateurs bayésiens du maximum a posteriori (MAP).

Pour les concentrations plasmatiques d'AMP, il s'agit par exemple de réaliser des prélèvements à T_{20min} T_{1h} et T_{3h} [9,21]. Pour l'évérolimus dans le sang total, des modèles d'estimation de l'AUC ont été proposés avec les temps T_0, T_{1h} et T_{3h} [22], ou uniquement T_0 [23], mais à chaque fois validés sur peu de patients, ou en utilisant des AUC de référence calculées à partir de peu de prélèvements.

Depuis 2005, le service de Pharmacologie, Toxicologie et Pharmacovigilance du CHU de Limoges met à disposition de tous les professionnels de santé qui suivent des patients transplantés, des modèles PopPK avec estimateurs bayésiens MAP pour les principaux médicaments immunosuppresseurs. Le système expert, disponible sur internet [24], est appelé « ABIS » pour Adaptation Bayésienne des ImmunoSuppresseurs. Cet outil permet d'estimer les AUC des immunosuppresseurs utilisés après une greffe rénale [14], hépatique [25], cardiaque [26,27], ou pulmonaire [27], ou pour des maladies auto-immunes [28], à partir de 3 concentrations sanguines et quelques informations sur le patient : organe greffé, âge, traitement immunosuppresseur associé, délai post-greffe (Figure 3). Chaque requête est

validée par un (ou plusieurs) pharmacologue expérimenté qui choisit le modèle le plus approprié, propose un ajustement de la posologie, et rend une conclusion en fonction du contexte de la demande. En France, c'est près de la moitié des patients transplantés rénaux qui ont pu bénéficier de ce suivi personnalisé [29].

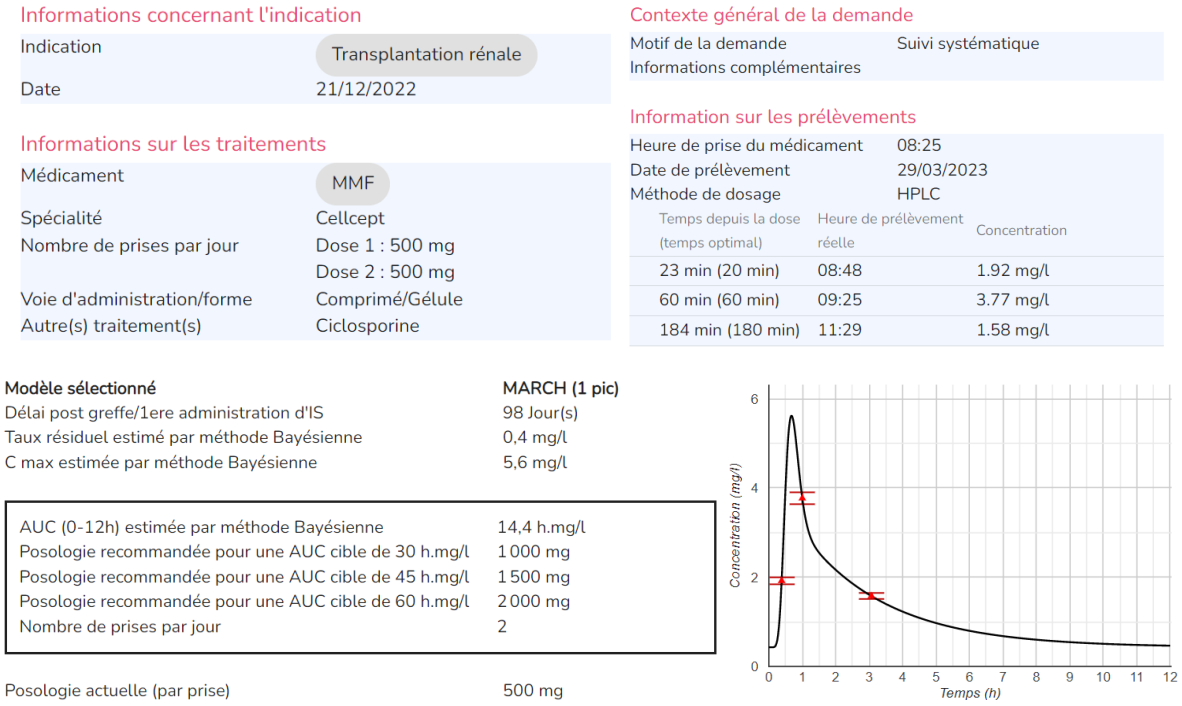


Figure 3 : Site internet d'Adaptation Bayésienne des ImmunoSuppresseurs (ABIS) : exemple de demande d'adaption individuelle de la dose de mycophénolate mofétil (Cellcept®)

AUC, aire sous la courbe des concentrations en fonction du temps ; HPLC, chromatographie en phase liquide à haute performance ; MARCH, acronyme du nom du modèle utilisé pour l'estimation ; MMF, mycophénolate mofétil.

Source : abis.chu-limoges.fr [24]

I.2. Les effets

Le *succès* de la greffe allogénique d'un organe dépend de l'immunosuppression thérapeutique. Les médicaments immunosuppresseurs ont toutefois des *effets indésirables*. Entre autres, ils augmentent le risque de développer des infections et des cancers. Le médecin doit être capable d'évaluer la balance bénéfices/risques propre à chaque situation, pour prendre les bonnes décisions avec le patient.

Nous nous concentrerons ici sur les effets attendus dans le cadre de la transplantation *rénale*.

I.2.1. Les bénéfices

Savoir *définir* les bénéfices thérapeutiques est primordial pour pouvoir comparer les stratégies thérapeutiques, proposer un suivi approprié, et optimiser la prise en charge de tout patient receveur, et ce de manière reproductible.

Dans le contexte de la transplantation, ces bénéfices « pharmacologiques » sont non seulement multiples, mais ils dépendent également des caractéristiques clinico-biologiques initiales du patient. Ce sont ces caractéristiques qui sont à l'origine de profils de risques *individuels*. Ils sont propres à chaque transplantation, puisqu'un même patient peut être greffé plusieurs fois avec un même organe, ou avec plusieurs organes.

Le principal bénéfice attendu est le maintien d'une *fonction rénale* permettant d'assurer la filtration du sang [30]. Elle est en générale estimée à partir de la concentration sérique de créatinine (MDRD [31], CKD-EPI [32]), en tenant compte ou non de la protéinurie. Ainsi en 2016, l'agence européenne du médicament a retenu comme pertinents les critères d'évaluation suivants : la fonction rénale à différentes étapes post transplantation, 6 et 12 mois, et l'incidence de la protéinurie (ou aggravation) [33]. La fonction du rein greffé dépend de la fonction initiale du greffon après transplantation et de nombreux facteurs de risque pour le greffon, en particulier le risque de rejet aigu mais aussi le risque infectieux, vasculaire, néphrotoxique, etc. Il s'agit donc de retarder les re-transplantations ou les *retours en dialyse*. L'échec est défini par la perte de la fonction rénale du greffon et la fin de la période d'autonomie, 1 an au moins après la transplantation [34]. En moyenne, la durée de vie d'un rein greffé est de 10 à 15 ans avant de devenir non fonctionnel. Néanmoins, ces résultats varient de manière importante entre les patients transplantés. Ils peuvent être en partie prévisibles en fonction de certaines caractéristiques individuelles : fonction rénale à différents moments de la première année post transplantation [35–37] (peu prédictif si considéré seul [38]), âge du donneur [39], immunisation pré ou post transplantation [40,41], voire de manière plus invasive d'après les résultats des premières biopsies [42,43]. Ainsi, de nombreux scores ont été développés pour tenter de prédire avec justesse la durée de vie à long terme d'un rein greffé [44–49]. Le « Score ajustable pour la prédiction de l'échec de la greffe » (AdGFS) est l'un d'entre eux [50] et a l'avantage de ne nécessiter que 2 à 7 variables (Figure 4), dans un arbre de décision facile à comprendre.

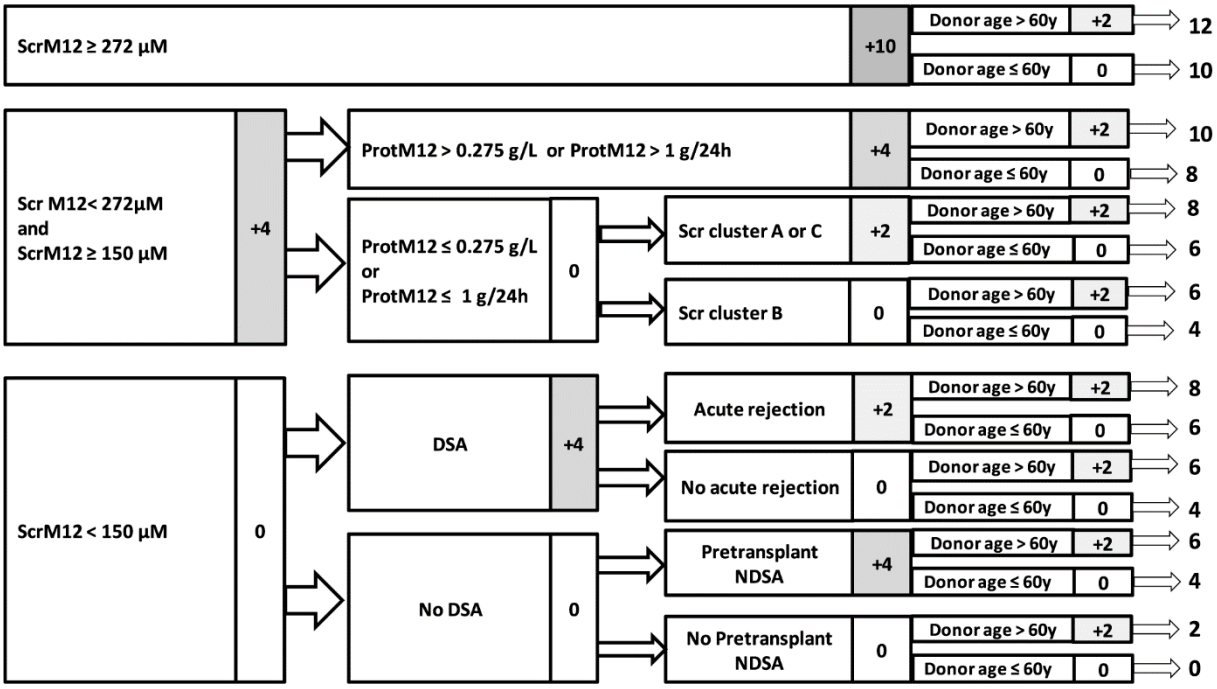


Figure 4 : Score ajustable pour la prédiction de l'échec de la greffe (adjustable graft failure score, AdGFS)

dnDSA, anticorps anti-HLA spécifique du donneur de novo ; NDSA, anticorps anti-HLA non spécifique du donneur ; ProtM12, protéinurie à M12 ; Scr, créatininémie ; ScrM12, créatininémie à M12.

Source : Prémaud et al. [57]

En fonction des branches, les informations utilisées sont : l'âge du donneur ; la présence d'anticorps anti-HLA non spécifique du donneur (NDSA) avant la transplantation ; le profil des concentrations sériques de créatinine durant la première année post transplantation, à partir des valeurs à 1 mois, 3 mois, 6 mois et 12 mois (Annexe 2) ; la protéinurie à 12 mois post transplantation ; l'apparition d'anticorps anti-HLA du donneur (DSA) ; la survenue d'un premier épisode de rejet aigu (rejet médié par les anticorps ou par les cellules T). Par exemple, un patient avec un score plutôt bas (ex. AdGFS ≤ 2) a une probabilité de survie du greffon jusqu'à 10 ans après la greffe d'environ 95 % (soit une probabilité de perte du greffon de 5%). La présence de facteurs de risques supplémentaires, comme l'apparition de DSA au cours du suivi, augmente la valeur du score (ex. AdGFS ≥ 6) et conduit à une probabilité de perte du greffon de 65 % à 8 ans et de 84 % à 10 ans post-transplantation.

Un autre bénéfice attendu de l'immunosuppression, cette fois-ci immédiat, est la *prévention de survenue de rejets aigus*. Pour le rein, le diagnostic de rejet est actuellement retenu (ou écarté) grâce à l'examen anatomopathologique d'une biopsie du greffon (gold standard). Une définition précise est proposée par la classification de Banff, par l'utilisation de critères histologiques avec des seuils.

Cette classification histologique a été établie pour la première fois en 1991 après une réunion à Banff, au Canada [51], et a depuis fait l'objet de mises à jour régulières : 1997 [52], 2001 [53], 2003 [54], 2005 [55], 2007 [56], 2013 [57], 2015 [58], 2019 [59]. Ces changements

fréquents (Figure 5 et Figure 6), les nombreuses règles et exceptions (Annexe 1), et la nécessité de prendre en compte les données clinico-biologiques du patient (non incluses dans la classification de Banff), font de cette classification un outil difficile à manipuler pour les anatomopathologistes et néphrologues.

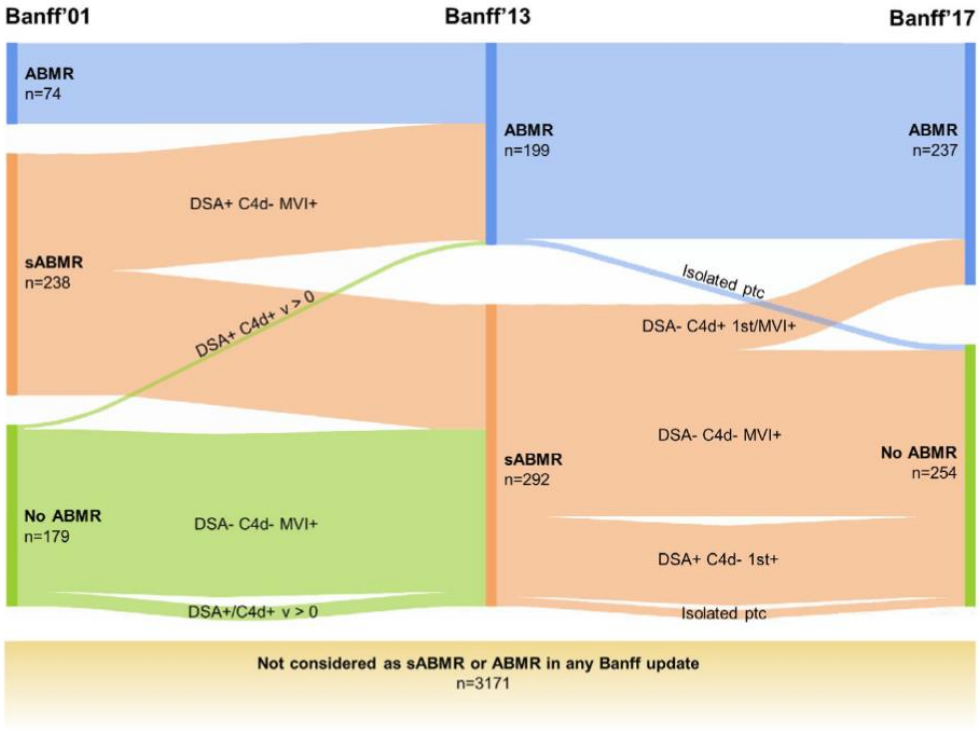


Figure 5 : Changements dans la définition du rejet médié par les anticorps (ABMR) de 2001 à 2017

Les diagnostics possibles sont : ABMR (bleu), ABMR suspecté (orange), absence d'ABMR (vert et jaune).

sABMR, rejet médié par les anticorps suspecté ; C4d, produit inactif issu de la dégradation catalytique du complément C4 ; DSA, anticorps anti-HLA du donneur ; MVI, invasion microvasculaire ; ptc, score de Banff de capillarite, présence de cellules dans les capillaires péritubulaires.

Source : Callemeyn et al. [60]

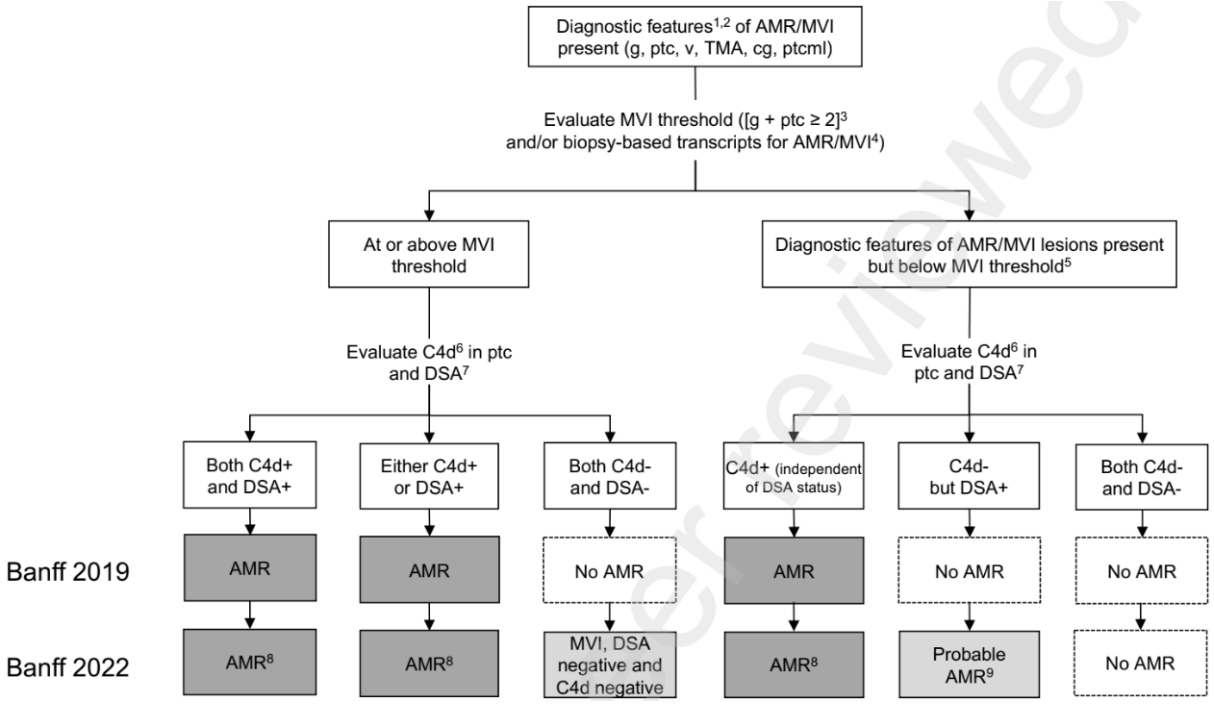


Figure 6 : Changements dans la définition du rejet médié par les anticorps (AMR) de 2019 à 2022

C4d, produit inactif issu de la dégradation catalytique du complément C4 ; cg, score de Banff pour les doubles contours sur la membrane basale glomérulaire ; DSA, anticorps anti-HLA du donneur ; g, score de Banff de glomérulite ; MVI, invasion microvasculaire ; ptc, score de Banff de capillarite, présence de cellules dans les capillaires péritubulaires ; ptcml, score de Banff pour la multilamellation de la membrane basale péritubulaire ; TMA, microangiopathie thrombotique ; v, score de Banff d'artérite intimale.

Source : Naesens et al., article soumis en cours d'examen par les pairs (The Banff 2022 Kidney Meeting Report: Re-Appraisal of Microvascular Inflammation and the Role of Biopsy-Based Transcript Diagnostics)

Les autres bénéfices attendus sont bien sûr ceux rapportés par le patient [61,62]. Ils incluent : la qualité de vie liée à la santé [63] ; les fonctions motrices ; la capacité à travailler ; l'observance du traitement ; et les symptômes spécifiques liés à la maladie ou à son traitement, par exemple, la douleur, la fatigue, les effets indésirables des médicaments. Ces derniers sont développés dans la partie qui suit, qui traite des « risques ».

I.2.2. Les risques

Les risques des traitements immunosuppresseurs peuvent être séparés en deux groupes : ceux liés à l'immunosuppression, comprenant les infections opportunistes et un risque accru de cancer ; et les autres effets indésirables, spécifiques à la classe médicamenteuse ou à la molécule (Figure 7).

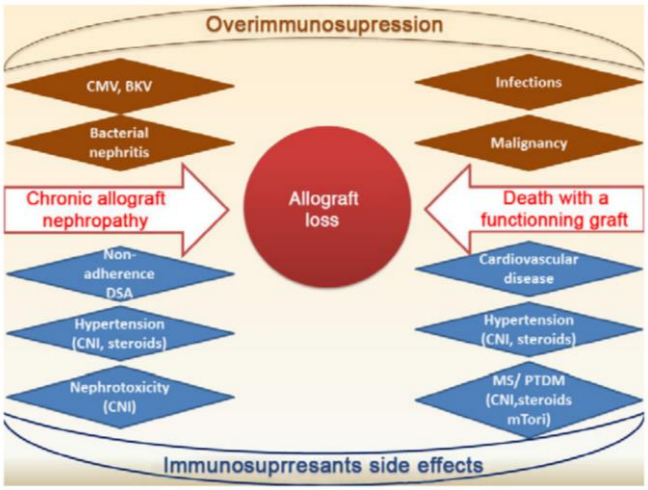


Figure 7 : Schéma des risques de l'immunosuppression (en marron), et des autres risques d'effets indésirables (en bleu)

BKV, virus BK ; CMV, cytomégalovirus ; CNI, inhibiteurs de la calcineurine ; DSA, anticorps anti-HLA du donneur ; MS, syndrome métabolique ; mTori, inhibiteurs de la mammalian target of rapamycin ; PTDM, diabète post-transplantation.

Source : Bamoulid et al. [64]

Les effets indésirables communs les plus fréquemment rapportés sont digestifs [65] : nausées, douleurs abdominales, et anorexie. Cependant, à forte dose, les immunosuppresseurs peuvent entraîner des effets indésirables plus graves [66] : les infections, par exemple par le cytomégalovirus [67] qui est le pathogène le plus fréquent, ou le virus BK [68–72] ; les syndromes lymphoprolifératifs liés à la réactivation de certains virus, comme le virus d'Epstein-Barr [73] ; et les cancers, dont les plus fréquents sont cutanés [74].

D'autres effets indésirables sont spécifiques à la molécule. Pour l'AMP, il s'agit d'hématotoxicité [75] : leucopénie, thrombopénie, anémie. Pour les CNI, ils induisent à long terme :

- une néphrotoxicité [76] : (1) aiguë réversible due à la vasoconstriction des artérioles afférentes qui entraîne une chute du débit de filtration glomérulaire, et (2) chronique irréversible et dose-dépendante par hyalinose artériolaire entraînant des lésions ischémiques (Figure 8) ;
- une neurotoxicité [77,78] : tremblements, céphalées, paresthésies, convulsions ;
- et une toxicité cardiovasculaire [79] : hypertension artérielle, tachycardie.

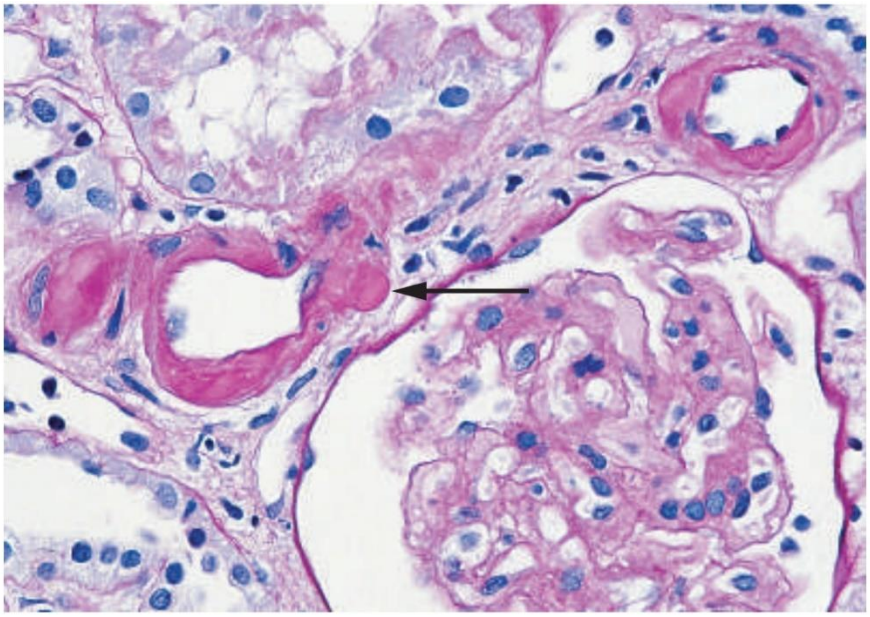


Figure 8 : Néphrotoxicité chronique induite par les inhibiteurs de calcineurine

Hyalinose au niveau de l'artériole afférente (flèche), correspondant au score ah de la classification de Banff. Coloration par l'acide périodique et réactif de Schiff (PAS) ; ×400.

Source : Liptak et al. [80]

La ciclosporine entraîne parfois une hyperplasie gingivale, un hirsutisme, de l'acné [78]. Le tacrolimus, lui, favorise le diabète post transplantation [78] (New Onset Diabetes After Transplantation, ou NODAT), en combinaison avec des facteurs préexistants propres au patient.

Concernant les inhibiteurs de mTOR, ils peuvent entraîner des hyperlipidémies [81,82] ou des thrombopénies [83], mais aussi des pneumopathies graves, retards de cicatrisation, aphtes et autres atteintes cutanéomuqueuses invalidantes.

Les immunosuppresseurs sont donc essentiels à la survie du patient en lui permettant de garder un greffon fonctionnel. Mais leur utilisation *à long terme* peut entraîner des effets indésirables graves, précipitant dans le pire des cas le décès (Figure 7 et Figure 9).

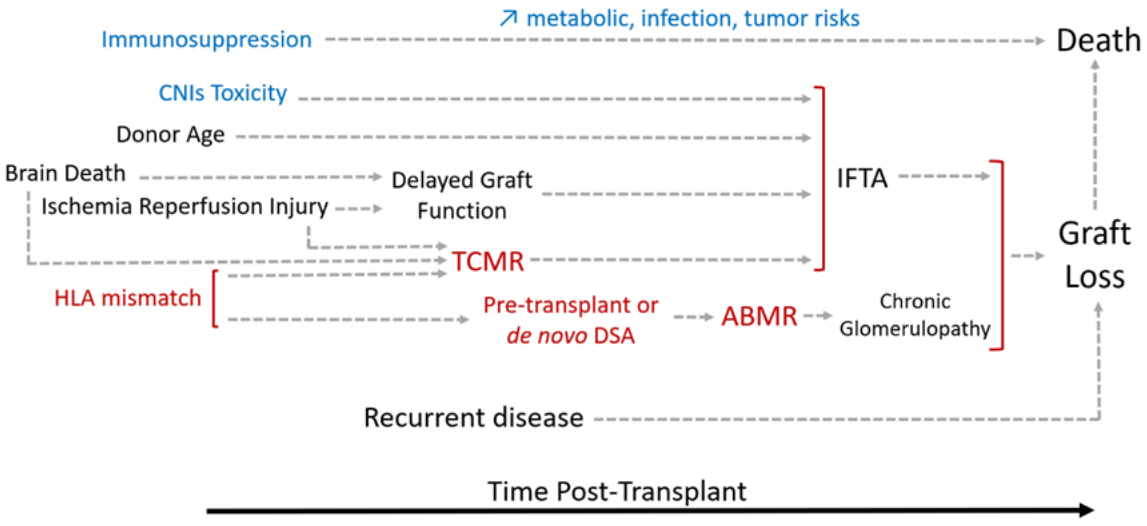


Figure 9 : Trajectoires possibles après une transplantation jusqu'à la perte du greffon ou le décès.

ABMR, rejet médié par les anticorps ; CNI, inhibiteurs de la calcineurine ; DSA, anticorps anti-HLA du donneur ; HLA, antigènes des leucocytes humains ; TCMR, rejet médié par les cellules T.

Source : figure modifiée à partir de la publication de Wiebe et al. [84]

Il peut également arriver que les effets indésirables conduisent à une baisse de la posologie ou un arrêt provisoire du traitement médicamenteux, ce qui favorise la survenue de rejets aigus et peut contribuer à long terme à la perte du greffon.

Il faut alors mettre en balance les bénéfices escomptés et les risques connus d'un traitement. Cette balance peut être évaluée lors des études cliniques qui étudient des *populations*, mais elle doit également être considérée à l'échelle *individuelle*, en fonction du terrain (âge, sexe, comorbidités) et des facteurs de risque individuels.

I.2.3. La balance bénéfices-risques

Dans certains cas, les risques ne peuvent pas être mesurés de la même manière que les bénéfices, et cela rend plus difficile la modélisation mathématique de cette balance.

En attendant de tels outils mathématiques, le clinicien doit se référer à la littérature afin d'arbitrer la prise en charge pour chaque patient.

Par exemple, le tacrolimus est considéré comme un CNI plus puissant que la ciclosporine [78], qui a été découverte en premier. Cependant, l'utilisation du tacrolimus entraîne des déséquilibres glucidiques chez certains patients à risque, et peut provoquer la survenue de diabète post-transplantation [85]. Le changement de CNI pour la ciclosporine permet alors de résoudre ces troubles dans de nombreux cas [86].

Une autre étude, elle, a évalué deux combinaisons d'immunosuppresseurs en comparant en même temps les bénéfices (empêcher la survenue de rejet aigu, de perte du greffon, ou de décès) et les risques virologiques (infections à cytomégalovirus et au virus BK) [87]. Les deux associations d'immunosuppresseurs étaient : évérolimus associé à un CNI à dose réduite, versus MPA avec un CNI à dose standard. Les infections étaient moins fréquentes dans le premier groupe avec une survie du greffon semblable au bout de 12 mois (il faut noter toutefois que c'est un critère très peu discriminant puisque la survie est de 10 à 15 ans). Dans une étude similaire, c'était l'utilisation de bélatacept (à forte ou faible dose) qui était comparée avec la ciclosporine, en termes de perte du greffon et d'effets indésirables [88]. La survie du greffon et celle du patient étaient meilleures sous bélatacept, sans différence entre les deux doses utilisées. Le bélatacept entraînant plus de syndromes lymphoprolifératifs post-greffes, il est néanmoins contre-indiqué chez les patients séronégatifs au virus d'Epstein-Barr [89].

A ce jour, très peu de modèles permettent d'estimer le bénéfice, ou le risque (ou les deux à la fois) de stratégies thérapeutiques en fonction du profil de la transplantation. Quand cela est fait, les études se basent sur des suivis sur de courtes périodes par rapport à la durée moyenne de survie d'un greffon.

I.3. La relation exposition-effet

Pour éviter les réactions de rejet du greffon, tout en minimisant certains effets indésirables comme le risque infectieux, les concentrations sanguines des médicaments sont utilisées comme biomarqueurs du niveau d'immunosuppression.

Des études ont par exemple montré que les infections sont plus fréquentes pour les patients avec des concentrations résiduelles de CNI plus élevées [90], mais la prévention des rejets nécessite également une exposition plus importante [78,91]. Starling et al. ont montré que le nombre de rejets aigus (prouvés par biopsies) était significativement réduit au-delà d'un certain seuil de concentration résiduelle en évérolimus [92]. Cependant, si cette concentration était trop élevée, les signes de toxicité étaient significativement plus fréquents [93].

A ce jour, les essais qui étudient en la relation entre le bénéfice et le risque sont rares et concernent de petits échantillons de patients [94]. Le Meur et al. ont ainsi montré que le taux de survenue d'infections à herpès virus était associée à une AUC d'AMP plus élevée, comme la réduction significative du risque de rejet aigu [95].

II. L'intelligence artificielle

II.1. Généralités

L'Intelligence Artificielle (IA) regroupe des algorithmes plus ou moins évolués qui *imitent* l'action humaine. En général, l'ordinateur applique ces tâches à une vitesse bien supérieure à celle d'un être humain. Dans sa version la plus basique, la logique « *if… then… else…* » dans un programme en fait une IA, sans qu'elle soit forcément vraiment « intelligente » (Figure 10).

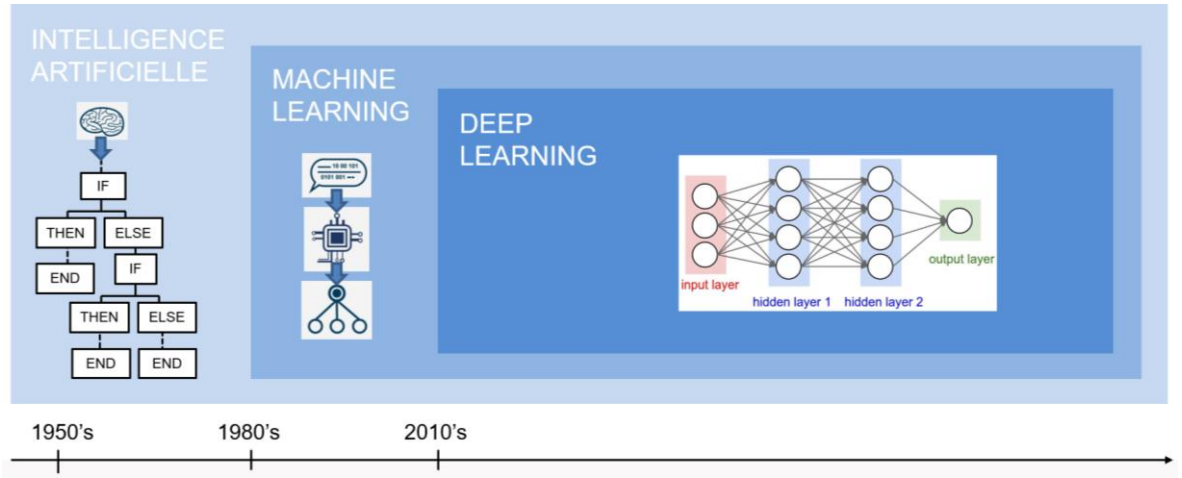


Figure 10 : Machine Learning et Deep Learning : deux parties constituantes de l'intelligence artificielle.

Le Machine Learning (ML) est un sous-domaine de l'intelligence artificielle. Comme son nom l'indique, dans le ML, l'ordinateur apprend des données qu'on lui fournit pour pouvoir les appliquer sur un nouveau cas. Pour ce faire, le ML va s'appuyer sur des algorithmes *statistiques* découverts il y a quelques décennies, mais récemment optimisés pour une utilisation plus rapide et efficace. C'est par exemple le cas du gradient boosting, découvert à la fin des années 1990 [96,97] et optimisé en 2016 dans la méthode appelée eXtreme Gradient Boosting (XGBoost) [98]. Les forêts aléatoires, quant à elles, ont été découvertes au milieu des années 1990 [99].

Le Deep Learning, quant à lui, est un sous-domaine du ML qui va utiliser des *réseaux de neurones artificiels*. Ces derniers sont capables de prendre en compte des données très complexes comme les couleurs de chaque pixel d'une image, pour parvenir à reconnaître ce qu'elle représente. C'est seulement à partir des années 2010 que le *Deep Learning* a connu un véritable essor grâce à l'accélération des calculs.

II.1.1. Les méthodes par arbres

Plusieurs méthodes de ML utilisent les *arbres de régression* pour faire des estimations quantitatives. Ils sont une adaptation pour les données continues d'une technique de classification appelée « arbre de décision » [100]. Dans les versions les plus simples des arbres de régression, les feuilles contiennent une valeur moyenne à approcher (Figure 11).

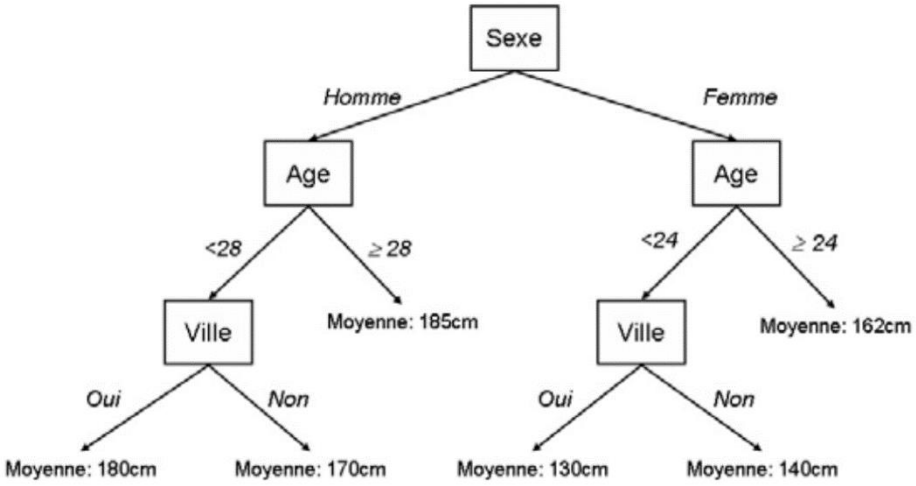


Figure 11 : Exemple d'arbre de régression

Dans cet exemple, la variable à approcher est la taille des individus en fonction de leur sexe, leur âge, et leur lieu d'habitation.

Source : figure modifiée à partir de la publication de Poli et Carrive [101]

Ainsi dans cet exemple, pour un individu donné dans une feuille, l'arbre de régression permet d'estimer la taille, avec une certaine erreur par rapport à la réalité. Pour accroître les performances du modèle, les méthodes de ML vont s'appuyer sur des techniques ensemblistes qui utilisent une combinaison de plusieurs centaines d'arbres (Figure 12).

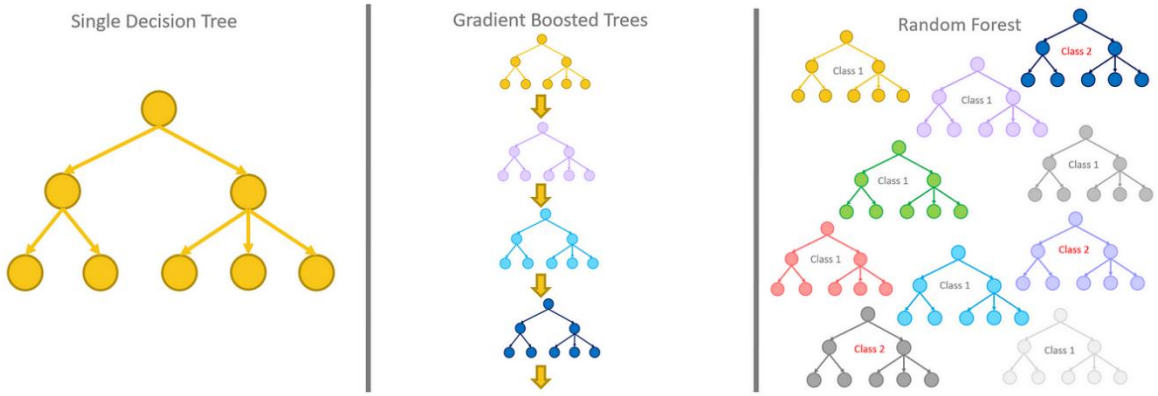


Figure 12 : Exemple de deux méthodes d'ensemble : le Boosting et le Bagging

Le Boosting (au milieu) est utilisé dans le gradient boosting, et le Bagging (à droite) est utilisé dans la forêt aléatoire.

Source : Silipo [102]

Dans la construction des modèles, le Boosting travaille de manière séquentielle. Il commence par construire un premier modèle (un premier arbre) qu'il va évaluer. A partir de cette mesure, chaque individu va être pondéré en fonction de la performance de la prédiction. L'objectif est de donner un poids plus important aux individus pour lesquels la valeur a été mal prédite, pour la construction du modèle suivant. Le fait de corriger les poids au fur et à mesure permet de mieux prédire les valeurs difficiles. Le deuxième arbre va ainsi essayer de corriger les éventuelles erreurs du premier arbre. Au final, plutôt que d'utiliser un seul arbre, plusieurs sont agrégés de manière séquentielle pour obtenir un seul résultat.

Un arbre de décision seul est susceptible d'entraîner du surapprentissage, c'est-à-dire des prédictions trop précises qui ne sont applicables qu'au jeu de données d'entraînement. C'est pourquoi la méthode de la forêt aléatoire, elle, consiste à agréger un grand nombre d'arbres *volontairement différents* [103]. Le caractère aléatoire du processus vient d'abord du fait que les arbres sont construits sur des échantillons différents : à partir de tirages avec remise du jeu d'apprentissage [104]. Puis un second niveau aléatoire est introduit à chaque étape de la construction des arbres : à chaque nœud, seul un sous-ensemble de variables est sélectionné aléatoirement pour choisir la coupure. La coupure la plus efficace n'est donc pas choisie parmi *toutes* les variables disponibles, mais parmi un sous-ensemble limité. A la fin, un système de vote est appliqué, où chaque arbre apporte une prédiction, et la valeur prédite finale est une moyenne de ces votes.

Il faut noter qu'en plus des coupures, les modèles sont caractérisés par des « hyperparamètres » qu'il faut optimiser pendant la phase d'apprentissage : nombre de variables disponibles pour faire la coupure la plus efficace à un nœud, profondeur maximale de l'arbre, nombre d'observations minimales dans un nœud...

II.1.2. Propriétés des méthodes de Machine Learning

Les modèles statistiques de ML présentés ici ont donc de nombreux avantages. Les variables explicatives peuvent être qualitatives (Figure 11, sexe homme ou femme, ville ou campagne) ou quantitatives (âge). Les modèles ne dépendent pas d'hypothèses sur la distribution des variables (comme la loi normale) [105], et pour certains, ils ne nécessitent pas de normalisation[2] des valeurs numériques (cas des méthodes qui utilisent des arbres [106]). Les arbres peuvent aussi bien être utilisés pour de la régression (estimation quantitative) que pour de la classification (arbres de décision) [100]. De plus, ces modèles sont capables de sélectionner les variables les plus efficaces parmi des centaines, et ce, même si elles sont colinéaires [107].

Pour chaque variable, des indices d'« importance » peuvent être calculés à l'échelle du *modèle*. Par exemple, si la variable choisie à un nœud est efficace et souvent choisie dans les nœuds, son « importance » mesurée sera grande, 1 étant le maximum (Figure 13) [103].

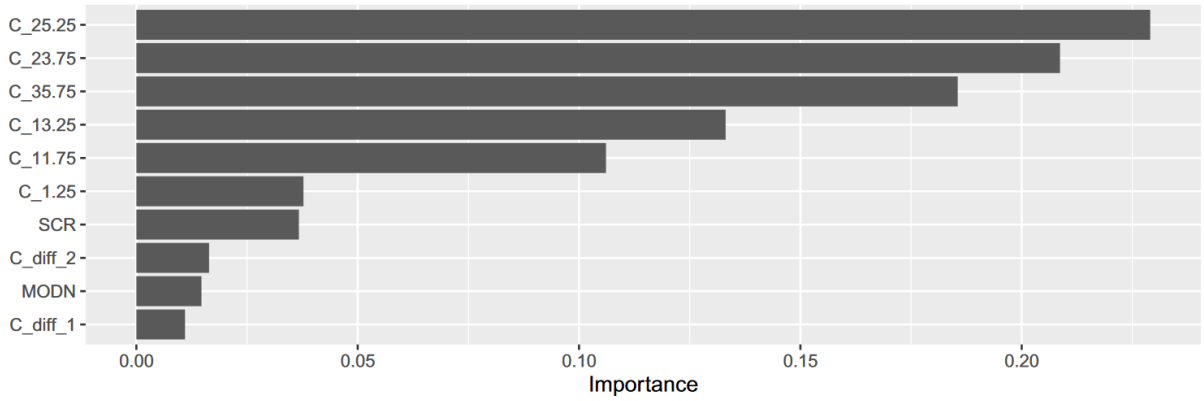


Figure 13 : « Importances » relatives des variables prédictives d'un modèle de Machine Learning pour l'estimation d'AUCs des concentrations sanguines de vancomycine

C_X, concentration mesurée au temps T_{Xh} ; MODN, type de population simulé (catégoriel) ; SCR, créatininémie, C_diff, différence entre la concentration résiduelle et la concentration au pic sérique.

Source : Bououda et al. [102]

Cependant, du fait du nombre important d'arbres utilisés dans chaque modèle, il est impossible de les représenter graphiquement. Cela contribue à l'effet *black box* reproché à ces méthodes. Pour mieux expliquer les prédictions, des techniques permettent de quantifier l'impact de chacune des variables sur la prédiction finale. C'est le cas de l'approche Shapley Additive exPlanations (SHAP) [108] illustrée ci-dessous, Figure 14.

[2] La normalisation « Min-Max » est une méthode qui met à l'échelle les données afin qu'elles soient bornées entre [0,1]. Elle ne doit pas être confondue avec la standardisation qui, elle, consiste à donner à une variable une moyenne égale à 0 et un écart-type de 1.

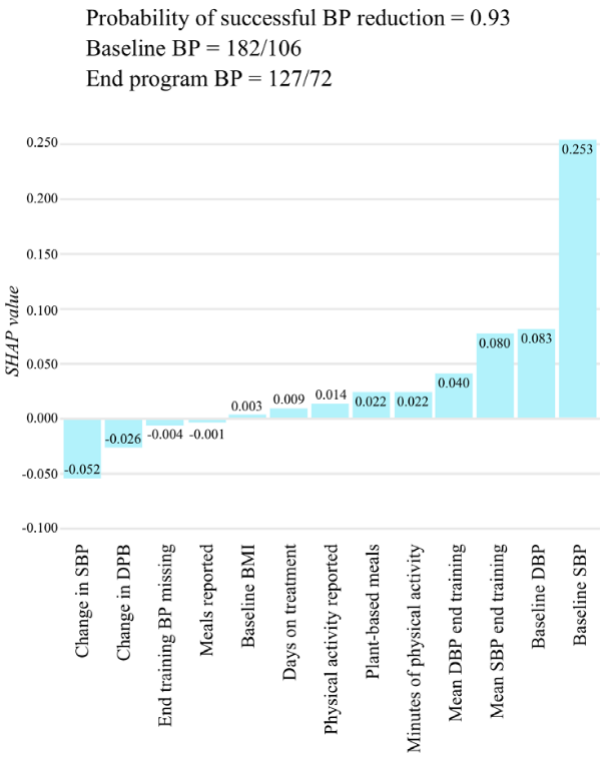


Figure 14 : Valeurs de Shapley des variables explicatives d'un individu donné

La valeur de Shapley représentée sur l'axe des ordonnées indique la quantité de contribution (positive ou négative) de la variable à la prédiction d'un seul individu. Dans cet exemple, le modèle était entraîné sur un petit échantillon de 135 adultes hypertendus qui recevaient des conseils sur une application. Il permettait de prédire une probabilité de diminution de la pression artérielle à 3 mois, en fonction des informations rapportées dans une application pendant le 1er mois. Dans cet exemple, elle était de 93%. Ceci est en partie expliqué par une pression artérielle systolique élevée de base dans ce cas précis, cet individu avait également indiqué prendre un certain nombre de repas avec des fruits et légumes, et passer du temps à pratiquer une activité physique.

BMI, indice de masse corporelle ; BP, pression artérielle ; DBP, pression artérielle diastolique ; SBP, pression artérielle systolique.

Source : Guthrie et al. [109]

Ces méthodes statistiques sont reconnues pour leurs performances [110]. En raison de tous les points forts précités, elles sont particulièrement intéressantes à utiliser en santé et en pharmacologie, où les données sont abondantes, complexes du fait de leur interdépendance, et difficiles à combiner.

II.2. L'apprentissage supervisé

L'apprentissage supervisé est la catégorie de ML la plus populaire. Il consiste à utiliser des jeux de données étiquetées, pour entraîner des algorithmes à classer des individus, ou prédire des résultats avec précision. La variable à prédire est donc connue lors de l'apprentissage et est appelée en général Y. Les autres variables qui vont être utilisées pour

construire le modèle sont appelées variables X (prédicteurs). Les méthodes citées dans la partie précédente Généralités sont des exemples d'apprentissage supervisé.

Cette catégorie est organisée en 2 sous-catégories. Quand Y est une variable qualitative (catégorielle), on parle de classification. Quand Y est quantitative, il s'agit de régression. (Figure 15).

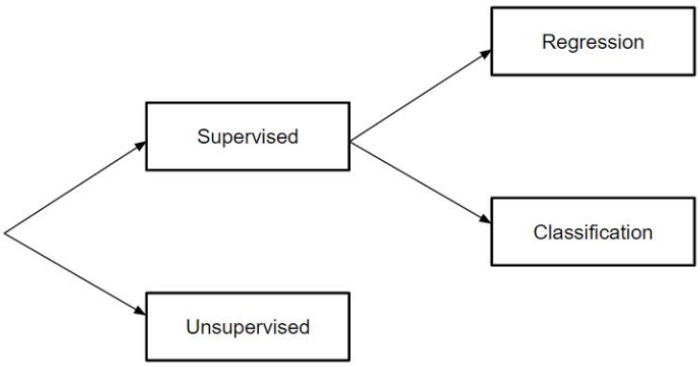


Figure 15 : Segmentation fondamentale des modèles de Machine Learning

Source : article en ligne [111]

II.2.1. Classification

Les algorithmes de classification utilisés en ML permettent de classer différents éléments d'un jeu de données en plusieurs catégories (qui sont connues pendant la phase d'apprentissage).

En 2019, Tomašev et al. (Google DeepMind) ont par exemple utilisé des réseaux de neurones récurrents pour prédire la survenue d'insuffisance rénale aiguë dans les 48h, dans une étude rétrospective incluant environ 700 000 patients adultes [112]. Leur modèle de Deep Learning prenait en compte plus de 300 prédicteurs issus de données d'hospitalisation dont : des données démographiques, des informations à l'admission, des constantes vitales, une sélection de médicaments prescrits, d'examens de laboratoires, et les diagnostics de maladies chroniques susceptibles de favoriser une insuffisance rénale aiguë. L'algorithme entraîné avec 80% des données a permis de prédire 56% de toutes les insuffisances rénales, et 90% des insuffisances rénales ayant nécessité une hémodialyse, avec un rapport de 2 faux positifs pour un vrai positif.

Dans un autre domaine, la reconnaissance d'image au microscope, ce sont des images de lames numériquement annotées qui servent de base d'entraînement. Des réseaux de neurones ont par exemple été entraînés à identifier les compartiments pertinents (par exemple, capillaires, tubules, vaisseaux et glomérules) nécessaires à l'évaluation des scores de Banff, pour le diagnostic du rejet [113–115]. Dans le futur, ils permettront très probablement d'avoir une estimation objective des aires atteintes par l'inflammation.

II.2.2. Régression

Les algorithmes de régression prédisent des valeurs *numériques* (connues pendant l'apprentissage) à partir des données d'un patient (prédicteurs).

En 2020, Woillard et al. ont ainsi entraîné un algorithme de ML, XGBoost, à estimer l'AUC de l'AMP [116]. Le modèle utilisait les prédicteurs suivants : concentrations plasmatiques d'AMP lors de 3 prélèvements sanguins effectués environ 20 min, 1h et 3h après la prise du médicament, temps de prélèvement réels, dose, indication du traitement, type d'immunosuppresseur associé, âge (et quelques autres variables issues de la transformation de celles précitées). La base d'apprentissage comprenait environ 13 000 $AUC_{0\text{-}12h}$ issues du site internet ABIS. A la fin, la validation externe sur des bases de données de profils complets de concentrations, a permis de mettre en évidence de meilleures performances de l'algorithme XGBoost par rapport à un estimateur bayésien MAP classique.

II.3. L'apprentissage non supervisé

Dans le cas de l'apprentissage non supervisé, des algorithmes de ML sont utilisés pour examiner et regrouper des jeux de données non labélisés, c'est-à-dire sans a priori sur le devenir du patient, le diagnostic final etc.

La technique d'apprentissage non supervisée la plus utilisée est le *clustering*.

Vaulet et al. ont par exemple appliqué cette méthode aux scores des lésions histologiques observées par les anatomopathologistes sur les biopsies de patients transplantés rénaux [117]. Une quinzaine de scores existent, semi-quantifiés de 0 à 3. Habituellement, la classification de Banff permet de classer les biopsies à partir de règles complexes, en fonction de combinaisons de scores choisies par consensus d'experts. Ici, l'algorithme des k-moyennes a été appliqué à la technique du clustering consensuel, pour regrouper de manière statistique les biopsies. Pendant la phase d'apprentissage, les auteurs ont légèrement influencé la création des clusters en fonction du devenir des greffons (apprentissage semi-supervisé par la survie des greffons), afin de créer des groupes pertinents sur le plan clinique et phénotypique. Une fois les groupes constitués, la méthode était appliquée à de nouvelles biopsies. Au final, le modèle a permis de constituer de nouvelles classes ayant des pronostics différents (Figure 16).

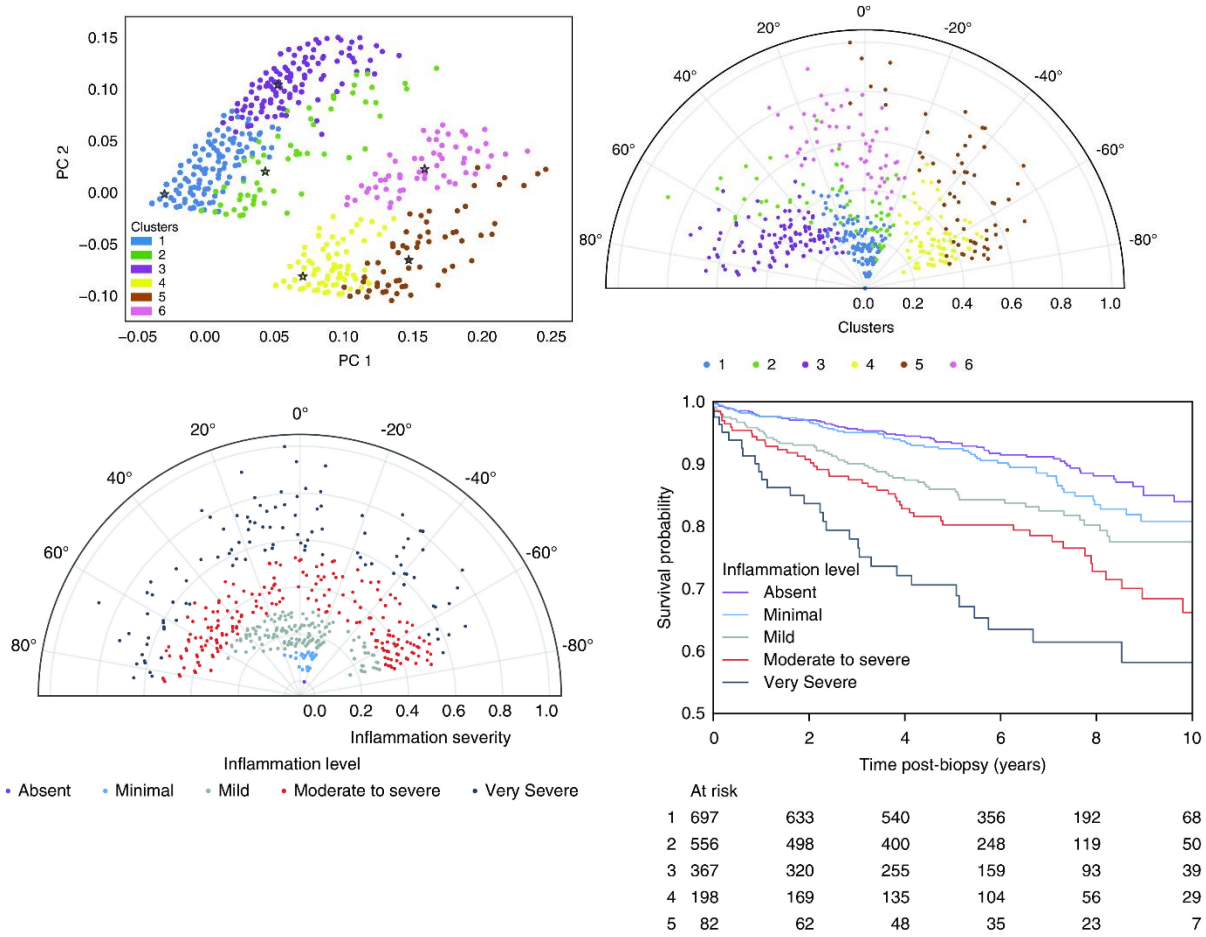


Figure 16 : Nouvelles classes de biopsies à partir des scores histologiques et de la survie du greffon

Les 6 clusters ont été construits par méthode semi-supervisée et représentés grâce à une analyse en composante principale (en haut à gauche). Sur le graphique polaire, les angles ont été obtenus à partir de la Composante Principale 2, et la somme des scores pondérés définit l'éloignement par rapport au centre (en haut à droite). Les nouveaux groupes d'inflammation ont été créés en fonction de cet éloignement (en bas à gauche). Les courbes de Kaplan-Meier correspondantes représentent la survie des 3510 greffons du jeu d'apprentissage (en bas à droite).

Source : Vaulet et al. [117]

Dans l'étude de Cippà et al. [118], les transcriptomes de biopsies ont été explorés à différents moments avec plusieurs méthodes non supervisées (classification hiérarchique, intégration de voisins stochastiques distribués en t, inférence de trajectoire…). Les gènes ont été regroupés en fonction de leurs similitudes d'expression. L'étude a permis de mettre en évidence des regroupements relatifs : à leur fonction (physiologie rénale, réponse à l'atteinte rénale, fibrose, immunité acquise), au moment de la biopsie (avant ou après transplantation), au type de don (donneurs vivants ou ceux décédés après arrêt cardiaque) et à la réponse phénotypique (DFGe, niveaux de fibrose quantifiés par le score de Banff 'ci'). La compréhension de ces mécanismes pourrait permettre de découvrir des prédicteurs utiles de la progression de lésions rénales après une transplantation.

Une autre technique d'apprentissage non supervisée très utilisée est la *réduction de dimensionnalité*. Il arrive qu'un jeu de données comporte un nombre de caractéristiques exceptionnellement élevé. La réduction de la dimensionnalité permet de réduire ce nombre sans compromettre l'intégrité des données.

C'est le cas des représentations graphiques de l'article de Reeve et al. où de nouvelles classes étaient construites à partir des phénotypes moléculaires de biopsies de patients transplantés rénaux [119]. Des puces à ADN ont permis d'étudier le transcriptome des 1208 biopsies et ainsi mesurer l'expression de plusieurs centaines de gènes au niveau des greffons. Les regroupements étaient là-encore non supervisés[3] (Figure 17). Six nouvelles classes étaient ainsi retenues, car elles ressemblaient de manière homogène à des formes de rejets connus. Une fois que chacune des biopsies était classée, les survies des nouveaux groupes étaient comparées. Les deux groupes « rejets médiés par les anticorps à un stade avancé » et « rejets médiés par les anticorps entièrement développés » avaient les plus mauvais pronostics.

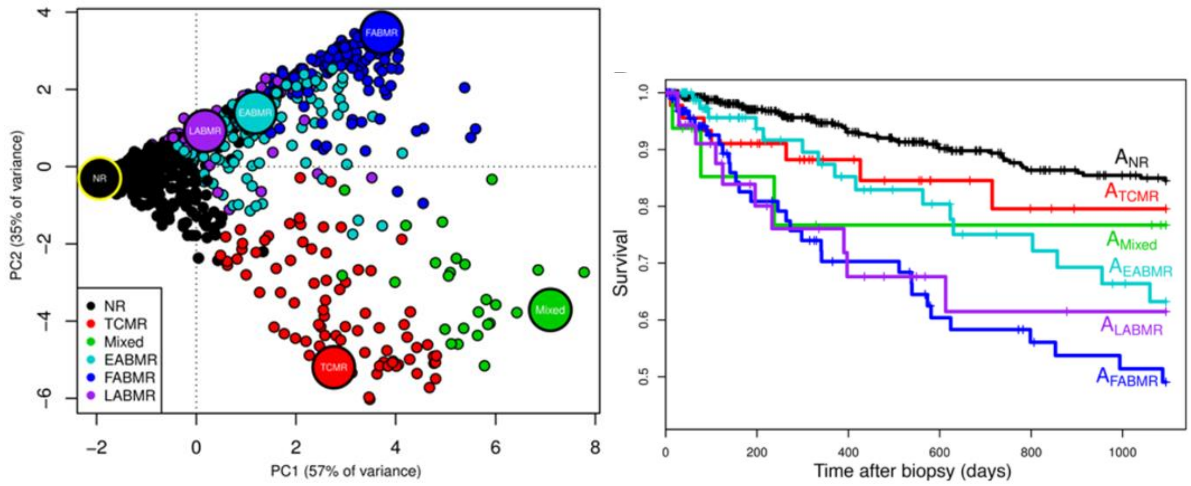


Figure 17 : Nouvelles classes de biopsies à partir des phénotypes moléculaires

Analyse en composante principale permettant de représenter toutes les biopsies entre elles en fonction de leurs caractéristiques en seulement deux dimensions (à gauche). Kaplan-Meier des différents clusters dont les noms ont été inspirés des diagnostics existants de la classification de Banff (à droite).

A, archetype (cluster) ; EABMR, rejet précoce médié par les anticorps ; FABMR, rejet médié par les anticorps entièrement développé ; LABMR, rejet médié par les anticorps à un stade avancé ; Mixed, rejet mixte médié par les anticorps et les cellules T ; NR, absence de rejet ; TCMR, rejet médié par les cellules T.

Source : Reeve et al. [119]

[3] Plus précisément, les expressions génétiques de chaque biopsie permettaient d'attribuer de manière supervisée des scores compris entre [0,1] pour la prédiction des 7 diagnostics : ABMR, TCMR et critères g, ptc, cg, i et t. Ces 7 scores pour chaque biopsie étaient alors utilisés pour faire l'analyse en composante principale, et les regroupements non supervisés par analyse archétypale.

II.4. Validation

II.4.1. Séparation des jeux de données

A l'origine de tous les modèles de ML, les données disponibles sont une ressource précieuse. Elles permettent dans un premier temps d'entraîner un algorithme, et ensuite, quand l'apprentissage est supervisé, de le tester (Figure 18). Les prédictions du modèle sont alors comparées aux observations réelles, qui servent de référence.

Figure 18 : Séparation des données en différents jeux

Les données peuvent être réparties de la manière suivante : 75 % pour l'apprentissage (bleu), 25 % pour le test (vert). Un jeu de données recueilli sur un autre site, par exemple issus de la pratique courante, peut servir à vérifier que le modèle est bien généralisable (rose).

Pour la création d'un modèle, les données d'entrée sont donc en général divisées en plusieurs jeux de données ayant chacun leur rôle.

Le jeu de données d'apprentissage ou d'entraînement va permettre de sélectionner les variables les plus pertinentes pour le modèle, et éventuellement d'écarter celles qui sont inutiles. Cette étape est importante pour proposer plus tard un modèle facile à utiliser en pratique courante. Il va également aider au choix des hyperparamètres (par exemple, profondeur maximale de l'arbre). Cette étape va aussi être utile pour donner un premier aperçu des performances du modèle imaginé. Mais la performance est peut-être juste l'effet d'une aubaine et d'un découpage particulièrement avantageux du jeu de données d'apprentissage. Pendant la phase d'apprentissage, la validation croisée ou cross-validation va donc permettre d'évaluer la performance d'un modèle de ML avec des résultats stables (Figure 19).

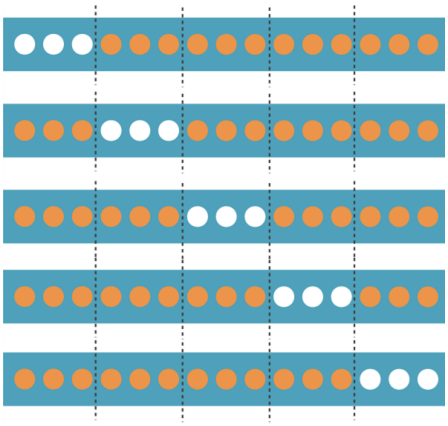


Figure 19 : Validation croisée à 5 folds pendant la phase d'apprentissage

Chaque point représente un individu dans la base de données. Chaque ligne représente un nouvel apprentissage. Cinq modèles vont être entraînés à partir de différents jeux d'apprentissage (points orange) et testés sur la partie ou fold restant (points blancs). Le jeu de données est donc divisé en 5 parties égales, et on utilise à chaque fois une partie comme jeu de test. La performance finale du modèle est calculée en faisant la moyenne des performances obtenues sur les 5 folds. En général, pour éviter des temps de calculs trop longs et inutiles, le nombre de folds est fixé à 5 ou 10.

Source : article en ligne [120]

A la toute fin de l'étape d'apprentissage, le jeu de données d'entraînement *entier* sera utilisé pour créer le modèle, avec les variables et les hyperparamètres choisis. Les paramètres optimaux du modèle sont alors « gelés » (par exemple, les nœuds utilisés dans un arbre de décision). L'étape qui suit est alors l'évaluation finale du modèle, appelée la phase de test. Le jeu de données de test fournira une évaluation impartiale du modèle ajusté sur le jeu de données d'apprentissage [121].

II.4.2. Surapprentissage

Certains algorithmes de ML qui recherchent dans les données d'apprentissage des relations empiriques entre la variable à prédire Y et les prédicteurs X ont tendance à surajuster les données (Figure 20). Ils risquent d'identifier et exploiter des relations apparentes dans les données d'apprentissage qui ne sont pas valables en général. Quand cela est possible, il est donc important de valider le modèle dans un jeu de données indépendant (extérieur) du jeu de données d'apprentissage.[4]

[4] Dans la littérature, les termes jeu de *test* et jeu de *validation* peuvent être inversés [122], le jeu de test se retrouvant alors être le jeu de données indépendant final qui sert à démontrer la généralisabilité d'un modèle.

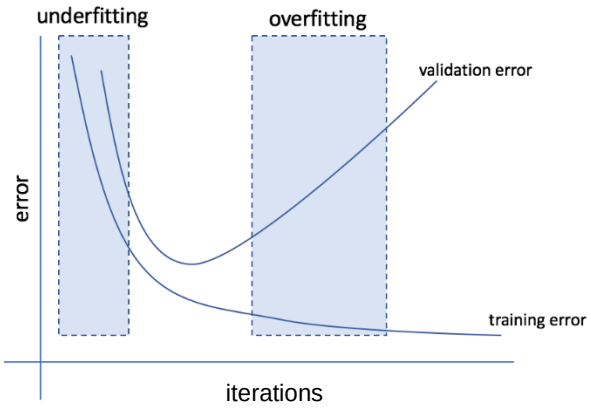


Figure 20 : Courbes des erreurs, et surapprentissage d'un modèle utilisant les données d'une seule source

Avoir un jeu de données suffisamment important et représentatif permet de diminuer l'erreur d'un modèle pendant l'apprentissage et en validation externe (à gauche). Si un modèle apprend des données issues d'une seule et même source, et si le nombre de données est très important, le modèle deviendra de plus en plus complexe jusqu'à expliquer les erreurs aléatoires du jeu, avec une diminution sans fin de l'erreur pendant l'entraînement (à droite, courbe du bas). Le modèle conviendra alors trop bien au jeu de données d'apprentissage au détriment de sa généralisabilité dans un jeu de validation externe (à droite, courbe du haut).

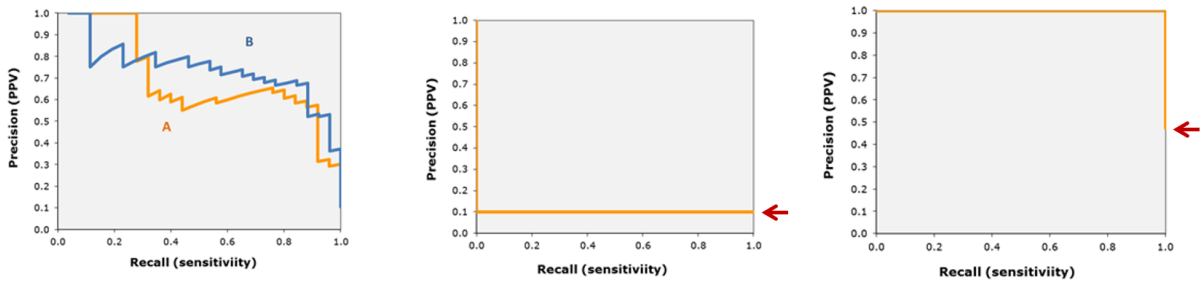
Source : figure modifiée à partir d'un article en ligne [123]

II.4.3. Courbes ROC et courbes Precision Recall

La plupart du temps, quand les performances d'un outil de classification sont évaluées, les données sont très déséquilibrées. Le nombre de cas négatifs (qui ne sont pas atteints de la maladie) est bien supérieur au nombre d'individus malades.

Lors de l'étape de validation d'un algorithme de ML, une courbe receiver operating characteristics (ROC) est souvent utilisée. Elle représente la sensibilité[5] en ordonnée et la spécificité[6] en abscisse pour différents seuils de positivité. Cependant, l'interprétation visuelle et les comparaisons de courbes ROC basées sur des jeux de données déséquilibrés peuvent être trompeuses. L'alternative à une courbe ROC est une courbe Precision Recall (PR) représentant le taux de vrais positifs (precision) en fonction de la sensibilité (recall) : Figure 21 et Figure 22.

[5] Sensibilité : capacité d'un score à rendre un résultat positif lorsque le patient est vraiment malade (rapport des vrais positifs sur tous les malades).
[6] Spécificité : capacité d'un score à rendre un résultat négatif lorsque le patient est vraiment indemne de la maladie (rapport des vrais négatifs sur tous les non malades).



example

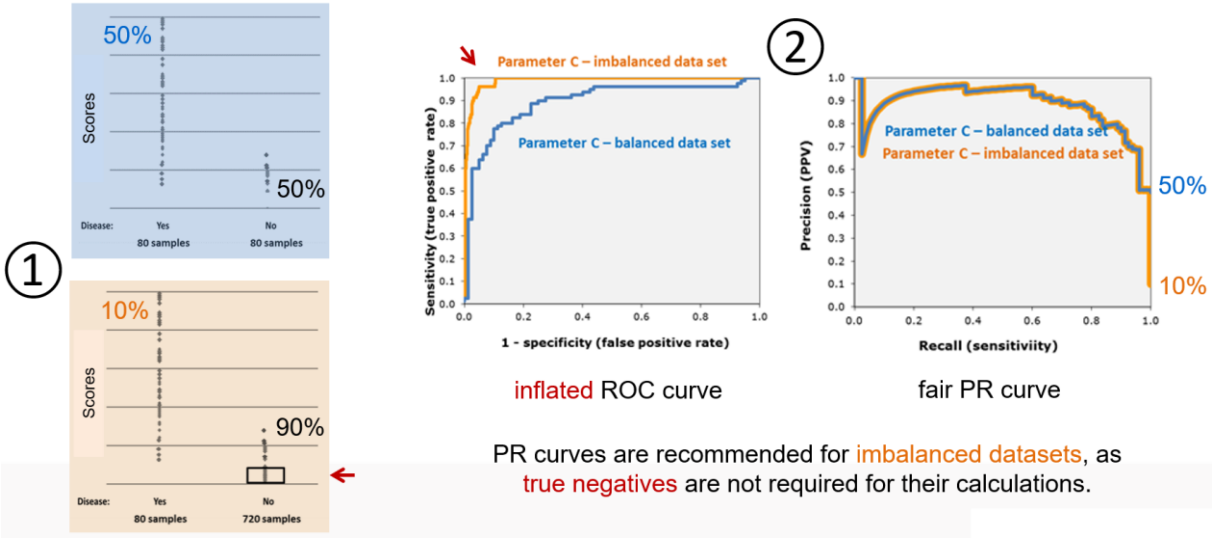
worst performance (disease with a prevalence of 10%):
AUC = 0.10

best performance (disease with a prevalence of 50%):
AUC = 1.00

Figure 21 : Exemples de courbes Precision Recall pour des scores de classification

Les performances de plusieurs exemples de modèles de classification sont représentées sur cette figure, en fonction de la prévalence (flèches rouges) : performances « classiques » pour un modèle A ou un modèle B (à gauche), pires performances possibles dans une population avec une prévalence de la maladie à 10 % (au milieu), maximum des performances atteignables dans une population avec une prévalence de la maladie à 50 % (à droite).

Source : figure modifiée à partir de la publication d'Ekelund [124]



inflated ROC curve

fair PR curve

PR curves are recommended for imbalanced datasets, as true negatives are not required for their calculations.

Figure 22 : Influence du nombre de cas négatifs sur la courbe ROC, quand dans le même temps la courbe Precision Recall n'est pas impactée

① Scores de différents individus malades (points gris à gauche) ou non malades (points gris à droite) pour une population équilibrée (bleue) et une population déséquilibrée avec une prévalence faible de 10 % (orange) ② Courbes ROC et courbes Precision Recall correspondant à une population équilibrée (bleue) et une population déséquilibrée avec une prévalence faible de 10 % (orange).

Source : figure modifiée à partir de la publication d'Ekelund [124]

En effet, quand le nombre de patients en bonne santé, souvent plus faciles à prédire, est très important, la courbe PR ne prend pas en compte le nombre de vrais négatifs (Figure 23).

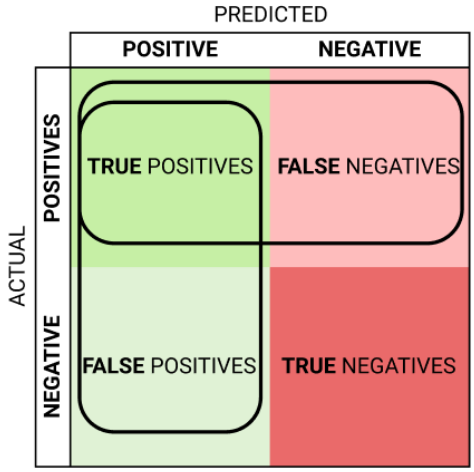


Figure 23 : Matrice de confusion pour l'évaluation d'un modèle de classification

Le nombre d'individus *malades* bien prédits est représenté en haut à gauche (vert intense). Le nombre d'individus *non malades* bien prédits est représenté en bas à droite (rouge intense). La courbe Precision Recall prend en compte le taux de vrais positifs (rectangle à gauche) en fonction de la sensibilité (rectangle en haut). Elle ne prend pas en compte les vrais négatifs (rouge intense) qui représentent souvent un groupe très important, et font gonfler artificiellement la courbe ROC.

Source : Shafi [125]

Dans une analyse de la littérature, Saito et al. ont observé une utilisation importante des courbes ROC pour des jeux de données déséquilibrés [126]. L'utilisation de courbes PR à la place des courbes ROC ferait changer les conclusions de nombreuses études. L'utilisation de courbes PR lors de la phase de validation de scores de classification, en supplément des courbes ROC, est donc hautement recommandée.

III. Objectif

Le succès d'une transplantation allogénique d'organe dépend de nombreux facteurs : qualité du greffon, conditions opératoires, modulation du système immunitaire, prise en charge médicale à long terme... Dans ce contexte, une médecine personnalisée bien conduite nécessitera de considérer les trois grandes dimensions que sont le patient, le greffon, et l'immunosuppression. A celles-ci, nous pouvons ajouter une quatrième dimension : le temps. Le défi d'une médecine de précision est donc de prendre en compte l'ensemble de ces variables interdépendantes pour faire les meilleurs choix possibles. Il faut en effet se servir des informations connues du patient (passé), pour proposer une prise en charge adaptée (présent), et anticiper les besoins à long terme (futur).

L'IA est l'outil idéal pour manier ces quatre dimensions. Les méthodes statistiques sont diverses, les algorithmes sont adaptables, et les logiciels sont accessibles. En outre, les ordinateurs ont depuis longtemps la capacité de gérer des puissances de calcul importantes. L'IA laisse donc présager une variété de possibilités et d'avantages pour l'exercice de la médecine en transplantation, mais elle soulève également des questions. Doit-elle poursuivre les mêmes objectifs que les outils statistiques traditionnels existant en pharmacologie clinique ? Est-elle utilisable malgré des jeux de données limités ? Peut-on améliorer des critères de jugement historiques ? Les modèles de prédiction à long terme sont-ils généralisables ?

L'objectif général de ce travail est de développer une utilisation appropriée de l'intelligence artificielle au service de la médecine de précision en transplantation.

Résultats

I. Mesurer l'apport des adaptations de posologie basées sur l'exposition à un médicament immunosuppresseur

I.1. Objectifs

Le site internet d'adaptation bayésienne des immunosuppresseurs (ABIS) [24] permet d'aider les médecins cliniciens et pharmacologues impliqués dans la prise en charge des patients transplantés. Il propose des ajustements de posologie grâce à l'estimation de l'exposition à des médicaments comme l'AMP, par le calcul de l'aire sous la courbe des concentrations.

Ce sont des estimateurs bayésiens MAP qui permettent de calculer ces AUC, à partir de seulement 3 prélèvements. Nous avons voulu évaluer ici l'intérêt de ces estimations d'AUC, pour justifier ou non la poursuite de cette stratégie avec des outils de ML. Les outils utilisés ici n'appartenaient donc pas au groupe des algorithmes de ML, mais aux modèles pharmacocinétiques et bayésiens plus classiques.

Dans cette étude, 4051 demandes d'ajustements de posologie ont été analysées de manière rétrospective, concernant 1051 patients transplantés rénaux âgés de 0-18 ans. Les estimations d'AUC avaient été réalisées sur le site ABIS à partir d'une stratégie de prélèvements en nombre limité. Trois mesures de concentration étaient suffisantes : $C_{20\,min}$, $C_{1\,h}$ et $C_{3\,h}$. Les informations nécessaires au choix du modèle le plus approprié étaient : l'indication du traitement, l'immunosuppresseur associé, la méthode de dosage utilisée, et le temps écoulé depuis la transplantation. Quand cela était nécessaire, une nouvelle dose étaient proposée pour viser une fenêtre thérapeutique de $AUC_{0\text{-}12\,h}$ = 30-60 mg.h/L comme recommandé [5].

Les objectifs spécifiques à cette première étude étaient : (i) de décrire l'exposition à l'AMP et plus précisément la proportion de patients ayant des AUC en-dehors de la fenêtre thérapeutique à la première utilisation du site internet ABIS ; (ii) d'étudier l'influence des différentes combinaisons d'immunosuppresseurs et des différentes périodes de l'enfance sur l'exposition à l'AMP ; (iii) d'évaluer l'efficacité des ajustements de posologie proposés aux cliniciens pour atteindre les AUC cibles.

I.2. Discussion

Cette étude justifie l'intérêt du suivi régulier des AUC d'AMP pour atteindre les cibles recommandées en transplantation rénale pédiatrique : $AUC_{0\text{-}12h}$ = 30-60 mg.h/L [5].

En effet, elle a permis de mettre en évidence le nombre important d'enfants sous-exposés lors de leur première estimation d'AUC (environ 50 %). La publication de Weber et al. [127] avait montré des résultats similaires pour un échantillon beaucoup plus petit de patients pédiatriques transplantés rénaux traités par MMF et ciclosporine (n = 25, comparés aux 749 demandes dans notre étude). Les enfants dont la dose était fixe et adaptée en fonction

de leur surface corporelle étaient sous-exposés à l'AMP pour 60 % d'entre eux. Une étude précédente de notre équipe chez des adultes, avait trouvé une proportion de « seulement » 25 % de patients sous exposés [14] (n = 13930), le plus souvent lorsqu'ils étaient traités par MMF et ciclosporine et au cours du premier mois qui suivait la transplantation. L'étude ne différenciait pas les « 1re estimations » des demandes d'estimation à la 2e visite où des ajustements de posologie avaient déjà pu être effectués.

A la 1re visite, l'exposition à l'acide mycophénolique était plus faible quand il était associé à la ciclosporine que lorsqu'il était associé au tacrolimus. Ceci est probablement dû à l'inhibition du cycle entéro-hépatique de l'acide mycophénolique par la ciclosporine (Figure 24). Les médecins transplanteurs ne sont peut-être pas assez bien informés de cette interaction médicamenteuse fréquente et de ses conséquences. Apparemment en tout cas, ils n'adaptaient pas la posologie a priori en fonction du CNI associé.

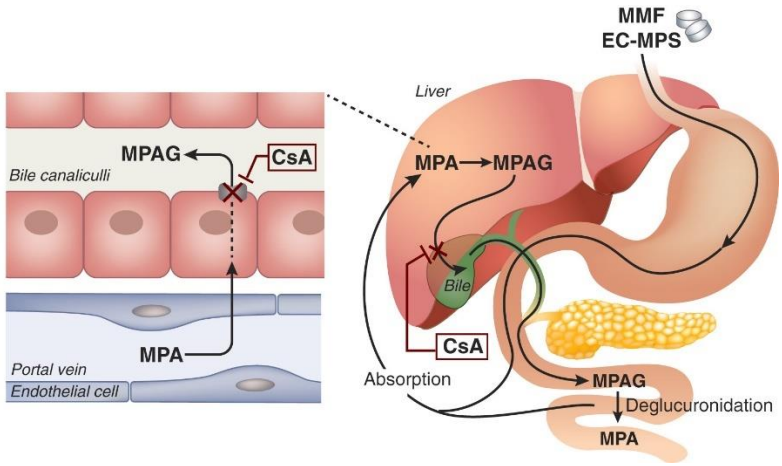


Figure 24 : Recirculation entéro-hépatique de l'acide mycophénolique

Après administration orale, le mycophénolate mofétil (MMF, ex Cellcept®) ou le mycophénolate de sodium à enrobage entérique (EC-MPS, ex Myfortic®)[7] est absorbé par l'intestin. Le métabolite actif, l'acide mycophénolique (MPA), est glucuronidé dans le foie et excrété dans la bile sous forme de glucuronide d'acide mycophénolique (MPAG, Annexe 3). Dans l'intestin, le MPAG est déglucuronidé par la flore intestinale, et le MPA est réabsorbé (recirculation entéro-hépatique). L'excrétion biliaire du MPAG de l'hépatocyte dans la bile est un processus actif, et la protéine de transport impliquée (multidrug resistance-associated protein-2 ou MRP2) est inhibée par la ciclosporine (CsA). Par conséquent, la recirculation est interrompue et l'exposition au MPA dans le plasma est réduite.

Source : van Gelder [128]

Dans l'étude présente, les ajustements proposés à la 1re visite de patients traités par MMF et ciclosporine étaient en moyenne de +306 mg (0-5 ans) et +344 mg (5-10 ans) pour viser une AUC de 45 mg.h/L. Quand c'était le tacrolimus qui était associé au MMF, les ajustements proposés étaient alors en moyenne de +371 mg (0-5 ans) et +173 mg (5-10 ans).

[7] Mycophénolate de sodium à enrobage entérique (EC-MPS) : comprimés fournissant la fraction active, l'acide mycophénolique, avec un enrobage entérique dans le but de diminuer les troubles digestifs.

Quand la posologie proposée à la 1re visite, était celle suivie (c'est-à-dire reçue) à la 2e visite, la proportion de patients dans les cibles d'AUC était plus importante que quand cet ajustement n'était pas respecté : 59 % versus 44 % (P = 0,002). Ces résultats ont été obtenus sans distinction entre les délais courts et longs entre les deux visites. Les résultats auraient peut-être été meilleurs en excluant les visites très éloignées. En effet, l'acide mycophénolique est connu pour présenter une variabilité intra-individuelle importante à long terme [14,129,130]. La variabilité des AUC autour de l'objectif de 45 mg.h/L était également significativement diminuée quand les posologies proposées étaient suivies. C'était particulièrement le cas 3 mois après la transplantation : P à 0,03 et 0,003 (Figure 2 de l'article). Quand les deux visites étaient espacées de plus de 1 an, et que la nouvelle posologie était respectée, cet effet était toujours présent mais moins important : P = 0,07.

Une des forces de cette étude est qu'elle reposait sur un nombre important d'AUC (n = 4051) dans cette indication précise qui est la transplantation rénale en pédiatrie. Les données étaient issues de nombreux centres indépendants ayant bénéficié du site internet ABIS (16 centres français différents avaient demandé plus de 30 ajustements de posologie sur la base de l'AUC).

Les AUC étaient calculées par méthode bayésienne grâce à une stratégie d'échantillonnage limité. Ce n'était pas des AUC « observées », calculées à partir de profils complets de concentrations. Cependant, ces modèles ont été *initialement validés* avec des profils riches en concentrations. Ces derniers permettaient de calculer des AUC de référence par méthode des trapèzes. En outre, le bruit (ou imprécision) lié à cette estimation ne correspond pas à un biais, puisqu'aucun facteur confondant n'a pu être identifié, et il n'a pas dissimulé les tendances significatives observées ici. En effet, les différents sous-groupes de cette étude étaient constitués de dizaines, voire de centaines de patients.

Dans notre étude, l'observation d'une proportion élevée d'AUC en-dehors de la cible a peut-être été aggravée par un biais de sélection. Pour certains patients, mesurer l'AUC de l'AMP était peut-être ponctuellement motivé par l'apparition d'effets indésirables observés biologiquement, ou la survenue de rejet aigu (suspecté par une dégradation de la fonction du greffon ou prouvé par biopsie). Les AUC ne représentaient donc pas de manière exhaustive et équilibrée toute la population pédiatrique traitée par acide mycophénolique dans l'indication d'une transplantation rénale.

Ce travail complète les études précédemment menées avec un nombre important d'adultes [14,15] et le niveau de preuve élevé des recommandations concernant la surveillance du MMF par l'AUC [5]. C'est une stratégie accessible en pratique clinique, et apparemment acceptée par les patients, les infirmières et les médecins, en raison de sa large utilisation dans les établissements de santé, essentiellement français.

Au final, cette étude, qui portait sur des modèles statistiques plus traditionnels, encourage la poursuite de cette activité d'estimation de l'AUC du MMF et l'amélioration des méthodes, par exemple grâce au ML. En 2021, Woillard et al. ont ainsi utilisé une technique de ML, appelée XGBoost, pour estimer les AUC d'AMP chez les patients transplantés adultes et pédiatriques [116]. Les performances en termes de RMSE étaient toujours meilleures que celles des estimateurs bayésiens MAP sur 4 jeux de données indépendants (n = 290).

I.3. Article 1

Mycophenolate mofetil dose adjustment in pediatric kidney transplant recipients

Ther Drug Monit 2023

Mycophenolate Mofetil Dose Adjustment in Pediatric Kidney Transplant Recipients

Marc Labriffe, MD,† Ludovic Micallef, PhD,* Jean-Baptiste Woillard, PharmD, PhD,*† Caroline Monchaud, PharmD, PhD,*† Franck Saint-Marcoux, PharmD, PhD,*† Jean Debord, MD,*† and Pierre Marquet, MD, PhD*†*

Background: The Immunosuppressant Bayesian Dose Adjustment web site aids clinicians and pharmacologists involved in the care of transplant recipients; it proposes dose adjustments based on the estimated area under the concentration–time curve (AUCs). Three concentrations ($T_{20\ min}$, $T_{1\ h}$, and $T_{3\ h}$) are sufficient to estimate mycophenolic acid (MPA) $AUC_{0–12\ h}$ in pediatric kidney transplant recipients. This study investigates mycophenolate mofetil (MMF) doses and MPA AUC values in pediatric kidney transplant recipients, and target exposure attainment when the proposed doses were followed, through a large-scale analysis of the data set collated since the inception of the Immunosuppressant Bayesian Dose Adjustment web site.

Methods: In this study, 4051 MMF dose adjustment requests, corresponding to 1051 patients aged 0–18 years, were retrospectively analyzed. AUC calculations were performed in the back office of the Immunosuppressant Bayesian Dose Adjustment using published Bayesian and population pharmacokinetic models.

Results: The first AUC request was posted >12 months posttransplantation for 41% of patients. Overall, only 50% had the first MPA $AUC_{0–12\ h}$ within the recommended 30–60 mg h/L range. When the proposed dose was not followed, the proportion of patients with an AUC in the therapeutic range for MMF with cyclosporine or tacrolimus at the subsequent request was lower (40% and 45%, respectively) than when it was followed (58% and 60%, respectively): $P = 0.08$ and 0.006, respectively. Furthermore, 3 months posttransplantation, the dispersion of AUC values was often lower at the second visit when the proposed doses were followed, namely, $P = 0.03$, 0.003, and 0.07 in the 4 months–1 year, and beyond 1 year with <6-month or >6-month periods between both visits, respectively.

Conclusions: Owing to extreme interindividual variability in MPA exposure, MMF dose adjustment is necessary; it is efficient at reducing such variability when based on MPA AUC.

Key Words: pediatrics, kidney transplantation, mycophenolate mofetil, mycophenolic acid, dose adjustment

(*Ther Drug Monit* 2023;00:1–8)

INTRODUCTION

Mycophenolate mofetil (MMF) is an immunosuppressant used to prevent kidney transplant rejection. Owing to the high interindividual variability in response to MMF therapy[1] and a narrow therapeutic window, therapeutic drug monitoring (TDM) has been recently recommended.[2] The TDM of the prodrug MMF is performed through the determination of its active metabolite, mycophenolic acid (MPA), and the MPA area under the concentration–time curve ($AUC_{0–12\ h}$) is the only exposure index associated with drug efficacy.[3–6] An $AUC_{0–12\ h}$ of 30–60 mg.h/L is targeted for both adults and pediatric transplant recipients.[2,7] The MPA AUC is measured using the trapezoidal rule; however, this requires several blood samples over 12 hours, which is time-consuming for the nurses or phlebotomists and the laboratory, in addition to being expensive. To estimate, rather than measure, MPA $AUC_{0–12\ h}$, adequate description of MPA plasma profiles was required. Therefore, one-compartment pharmacokinetic (PK) models with a double gamma distribution to describe the absorption phase,[8,9] which elicited accurate Bayesian estimation of the $AUC_{0–12\ h}$ using the limited sampling strategy $T_{20\ min}$, $T_{1\ h}$, and $T_{3\ h}$, were developed.[10] These PK models and Bayesian estimators are available and in use since 2005 on the Immunosuppressant Bayesian Dose Adjustment (ISBA) web site for assessing immunosuppressive drug exposure in many patients (https://abis.chu-limoges.fr).[11] Only 3 plasma MPA concentrations ($C_{20\ min}$, $C_{1\ h}$, and $C_{3\ h}$) and limited clinical data (MMF indication, combined immunosuppressant therapy, MPA assay technique, and time elapsed since transplantation) are used to estimate MPA PK profile and $AUC_{0–12\ h}$ and propose a more appropriate drug dose when necessary, to reach the 30–60 mg h/L target range. All results are examined and validated by experienced pharmacologists before being reported. The ISBA is currently used by nearly 100 transplantation centers worldwide, mainly for MMF and for tacrolimus (TAC), cyclosporine (CSA), everolimus, or sirolimus dose adjustments as well. The therapeutic indications are as follows: transplantation (kidney, liver, heart, lung, and bone marrow), nephrotic syndrome, lupus, and other autoimmune diseases.

Received for publication November 7, 2022; revision received December 9, 2022; accepted December 14, 2022.

From the *Pharmacology and Transplantation, INSERM U1248, Université de Limoges; and †Department of Pharmacology, Toxicology and Pharmacovigilance, CHU de Limoges, Limoges, France.

Conflicts of Interest and Source of funding: None declared.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.drug-monitoring.com).

Correspondence: Pierre Marquet, MD, PhD, Department of Pharmacology, Toxicology and Pharmacovigilance University Hospital of Limoges, CBRS, 2 Rue Bernard Descottes, 87000 Limoges, France (e-mail: pierre.marquet@unilim.fr).

The numerous routine requests for MMF dose adjustment posted on the ISBA for adult kidney transplant recipients (KTRs) were retrospectively analyzed 12 years ago.[12] This study focuses on pediatric KTRs and aims to (i) describe patient exposure to MPA and more precisely, the proportion of patients outside the therapeutic window, depending on age; (ii) examine the influence of the different immunosuppressant–MMF combinations on MPA exposure; and (iii) evaluate the efficiency of the dose adjustments on AUC target attainment proposed to clinicians.

MATERIALS AND METHODS

Patients

In this noninterventional study, MPA PK, exposure, and dose adjustment requests for pediatric KTRs addressed to the ISBA were retrospectively analyzed. This study was authorized by the French Committee for Informatics and Liberty, CNIL (authorization number: 1,619,537), and was exempted from IRB approval. The requests ranged from March 2010 to September 2019. Patients aged 0–18 years were included. All data were anonymized.

Database Description

After securely logging-in on the ISBA, local nephrologists (or clinical pharmacologists) fill in the following data: patient age, transplanted organ, date of transplantation, concomitant immunosuppressants, MMF dose, date, and exact time of blood sample collection, measured plasma MPA concentrations, and assay techniques. Thereafter, the web site proposes a selection of PK models and Bayesian estimators developed for the pediatric population to the expert clinical pharmacologists. The PK models combine a single-input or double-input (single or double gamma distributions) into a single compartment with linear elimination.[9,10] Similar to adult KTRs, $T_{20\ min}$, $T_{1\ h}$, and $T_{3\ h}$ are the limited sampling strategies best suited to the PK models and Bayesian estimators. Ten different models are available for pediatric KTRs receiving MMF, and the best model is chosen based on the following parameters: posttransplant period, combined immunosuppressants, patient age (if the patient is nearly 18 years old, adult models are tested as well), number of peaks compatible with the measured concentration–time points, and the compatibility of the modeled curves with these points. Based on the AUC estimate and previous dose, a range of doses are proposed to reach the boundaries and center of the 30–60 mg h/L range.

In this study, the following data for each patient and request were extracted: age, time elapsed posttransplantation, concomitant immunosuppressants, MMF dose, conventionally used MPA assay technique, estimated AUC, and the new dose proposed to reach the 45 mg h/L target.

Statistical Analyses

To observe the proportion of patients underexposed, overexposed, or adequately exposed to MPA and the potential influence of concomitant immunosuppressants, age, or posttransplantation period, the percentage of MPA AUCs <30 mg h/L or >60 mg h/L when MMF was administered with TAC, CSA, or another immunosuppressants were evaluated at 4 different posttransplantation periods: 1 month, 2–3 months, 4–12 months, or >12 months.

The χ^2 test was used to compare the percentages of MPA AUCs in the target interval at the first request between drug combinations.

The Student t test was used to compare the given and proposed doses at the first request in each subgroup of immunosuppressant combination, in the whole pediatric population and, subsequently, in the different age groups, namely, <5 years, 5–10 years, 10–15 years, and 15–18 years. Dose differences were tested globally across the different drug combinations using one-way analysis of variance.

To estimate the pertinence of the recommended doses, the percentage of AUCs outside the recommended range of 30–60 mg.h/L at request n + 1 and the distribution of AUCs regarding whether the recommended dose was followed were evaluated.

Moreover, the AUCs at the first visit were compared with those at the second visit using the Mann–Whitney–Wilcoxon test.

Furthermore, oral clearance (Cl/F) was studied and calculated using the following formula:

$$\frac{Cl}{F} = \frac{Dose}{AUC}$$

where Cl is clearance and F is the oral bioavailability factor.

All statistical analyses were performed using MedCalc version 20.113 (MedCalc Software bv, Ostend, Belgium), Microsoft Excel 2016 (Microsoft Corporation), and R version 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

In this study, 4051 dose adjustment requests, corresponding to 1051 pediatric patients from 51 transplantation centers worldwide, were analyzed. The median age (at the first request) was 13.1 years, and the mean MMF daily dose was 1000 mg. The concomitant immunosuppressants were TAC (74%), CSA (18.5%), or others (7.5%). Among all patients, 44%, 17%, and 39% had 3 or more, 2, and only 1 request posted, respectively. High-performance liquid chromatography was used to determine plasma MPA in approximately two-thirds of the requests. The AUC estimation and dose adjustment were first requested at 3, 4–12, and >12 months posttransplantation for 38%, 21%, and 41% of the cases, respectively.

Overall, only 50% of AUCs were in the target range at the first request. The corresponding distribution of AUCs for each drug combination is presented in Figure 1. The MPA AUCs were unbalanced over the first month of posttransplantation when MMF was combined with CSA; 39% and 3% of patients were underexposed and overexposed, respectively, compared with 30% and 20% of patients who were underexposed and overexposed, respectively, over the second and third months. When MMF was combined with TAC, the overall

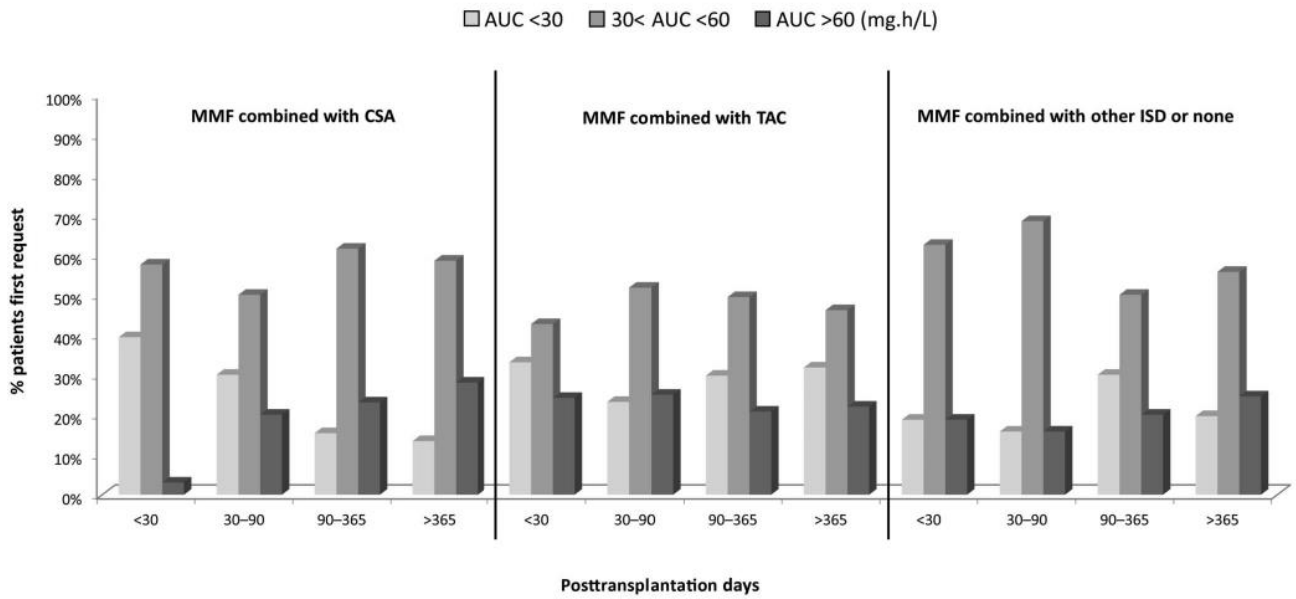


FIGURE 1. Distribution of MPA $AUC_{0-12\,h}$ at the first request for each drug combination. Underexposure (light gray), target range attained (medium gray), and overexposure (dark gray).

proportion of patients who were underexposed or overexposed was 53%, significantly larger than with CSA (43%, $P = 0.009$) or another immunosuppressants (42%, $P = 0.03$).

Table 1 lists the initial mean daily dose that was compared with the mean daily doses of MMF proposed to reach an AUC of 45 mg h/L. The difference was significant ($P < 0.05$) in almost every age group (proposed increase of approximately 200 mg/d). These differences were more pronounced for recipients aged <10 years, although the subgroups were smaller; for patients on CSA, an increase of 306 mg (0–5 years) and 344 mg (5–10 years) and for those on TAC, an increase of 371 mg (0–5 years) and 173 mg (5–10 years) were observed.

Subsequently, patients with 2 or more consecutive requests on the same drug combination were investigated (n = 472). When the doses recommended at the first request had not been followed, the proportion of patients with an AUC in the therapeutic range at the subsequent request (or second visit) was significantly lower than when it was followed: 44% versus 59%, ($P = 0.002$). For every drug combination (with CSA, TAC, or others), the proportion of patients with an AUC in the therapeutic range was always lower when the recommended dose was not followed (40%, 45%, and 43%, respectively) than when it was followed (58%, 60%, and 54%). It was significant for CSA and TAC (p=0.08 and 0.006, n = 108 and 344, respectively) but not for the group of other drug combinations, probably due to the smaller number (p=0.5, n = 20; Table 2). In addition, a potential center effect on the proportion of the dose recommendations that were followed was explored, which demonstrated no large variability (see **Figure, Supplemental Digital Content 1**, http://links.lww.com/TDM/A647, which illustrates the proportions of doses proposed and followed between various centers).

Figure 2 illustrates the distributions of MPA AUCs in patients who benefited from consecutive AUC measurements. Before 3 months posttransplantation, several pediatric KTRs were underexposed at the first or second request when the previously proposed dose was not followed. This proportion drastically decreased at the second visit when the previously proposed dose was followed; for the first and second–third months, median (interquartile range) AUC = 57 (37–82) and 51 (27–95) mg h/L versus 36 (14–60) and 37 (15–101) mg h/L when the proposed dose was not followed, respectively ($P < 0.001$ and $P = 0.32$, respectively). The P value in the second–third month was not significant probably because the number of cases was low (n = 14). Beyond 3 months posttransplantation, AUC dispersion had systematically decreased at the second request (minimum–maximum, interquartile range, or both), and this phenomenon seemed accentuated when the proposed dose was followed, for example, in the 4 months–1 year period, AUC SD (interquartile range) was 23 (21–76) mg h/L at the first visit versus 18 (25–70) mg h/L at the second visit when the proposed dose was followed ($P = 0.03$) and 24 (17–80) mg h/L when the proposed dose was not followed ($P = 0.85$). This was less significant when the second visit occurred more than 6 months after the dose recommendation.

Temporal evolutions of oral clearance are presented in Figure 3, considering different time scales, namely, for patient age and time posttransplantation (in the first year and over 15 years). The smooth line represents the trend of oral clearance and it moves through or close to the monthly or yearly medians (represented by black dots). Oral clearance increases beyond the age of 10 years. Posttransplantation, it reaches a nadir of 17.0 mL/min at approximately 4.5 months. A multilinear regression model of oral clearance reveals that age ($P < 0.001$) and time after transplantation ($P < 0.001$) significantly and independently influence Cl/F, without significant interaction between them ($P = 0.13$).

TABLE 1. Daily Doses of MMF Proposed to Reach AUC = 45 mg h/L, Sorted by Age and Combined ISD

		Combined Immunosuppressant			
		CSA	**TAC**	**Others**	***P* value**
Whole pediatric population					
*Requests		749	3001	301	
Daily dose of MMF (mg)	Median (min–max) at 1st request	1000 (160–3000)	1000 (80–3000)	1000 (200–3000)	
	Mean (±sd) at 1st request	1184 ± 546	1025 ± 501	1054 ± 568	**<0.001**
	Mean (±sd) proposed to reach AUC = 45 mg h/L	1416 ± 872	1220 ± 774	1133 ± 744	**0.002**
Dose given at 1st request vs. proposed dose (*P* value)		**<0.001**	**<0.001**	0.19	
			Age <5 yrs		
*Requests		90	166	29	
Daily dose of MMF (mg)	Median (min–max) at 1st request	760 (400–1800)	500 (80–960)	600 (320–1750)	
	Mean (±sd) at 1st request	848 ± 342	498 ± 205	639 ± 376	**<0.001**
	Mean (±sd) proposed to reach AUC = 45 mg h/L	1154 ± 600	842 ± 623	836 ± 471	0.06
Dose given at 1st request vs. proposed dose (*P* value)		**0.01**	**<0.001**	0.11	
			Age: 5–10 yrs		
*Requests		198	665	80	
Daily dose of MMF (mg)	Median (min–max) at 1st request	890 (160–1500)	760 (200–2000)	660 (240–1500)	
	Mean (±sd) at 1st request	844 ± 302	781 ± 349	738 ± 312	0.28
	Mean (±sd) proposed to reach AUC = 45 mg h/L	1215 ± 787	954 ± 585	759 ± 379	**0.001**
Dose given at 1st request vs. proposed dose (*P* value)		**<0.001**	**<0.001**	0.72	
			Age: 10–15 yrs		
*Requests		274	1036	99	
Daily dose of MMF (mg)	Median (min–max) at 1st request	1250 (500–3000)	1000 (160–2000)	1200 (320–2500)	
	Mean (±sd) at 1st request	1306 ± 526	1024 ± 421	1221 ± 562	**<0.001**
	Mean (±sd) proposed to reach AUC = 45 mg.h/L	1419 ± 753	1258 ± 723	1498 ± 888	**0.09**
Dose given at 1st request vs. proposed dose (*P* value)		0.19	**<0.001**	0.05	
			Age: 15–18 yrs		
*Requests		187	1134	93	
Daily dose of MMF (mg)	Median (min–max) at 1st request	1725 (220–3000)	1500 (200–3000)	1500 (200–3000)	
	Mean (±sd) at 1st request	1617 ± 523	1357 ± 498	1448 ± 562	**0.004**
	Mean (±sd) proposed to reach AUC = 45 mg h/L	1791 ± 1094	1484 ± 875	1305 ± 760	**0.04**
Dose given at 1st request vs. proposed dose (*P* value)		0.19	**0.03**	0.29	

Bold indicates statistical difference at $p < 0.05$.
*, No. of requests; ISD, immunosuppressant drug.

DISCUSSION

In this study, many of the first requests posted on the ISBA for MMF dose adjustment in pediatric KTRs were late (41% beyond the first year posttransplantation). Only 50% of all patients had an MPA $AUC_{0–12\ h}$ within the recommended range of 30–60 mg.h/L at the first request. The subsequent $AUC_{0–12\ h}$ was more often in the target range when the ISBA recommended MMF dose was followed (59% vs. 44%, $P = 0.002$).

TABLE 2. Efficiency of Dose Adjustment Recommendations (Restricted to Patients With Two or More Consecutive Requests and Same Drug Combination, n = 472)

	First Visit								
	MMF + CSA			MMF + TAC			MMF + Another ISD or None		
	No. of Patients (n)	Mean Daily Dose of MMF ± SD	Patients with 30 < AUC <60, in %	No. of Patients (n)	Mean Daily Dose of MMF ± SD	Patients with 30 < AUC <60, in %	No. of Patients (n)	Mean Daily Dose of MMF ± SD	Patients with 30 < AUC <60, in %
<5 yrs	20	871 ± 353	50	31	487 ± 210	42	2	460 ± 198	0
5–10 yrs	25	864 ± 274	44	83	804 ± 366	45	5	640 ± 267	60
10–15 yrs	35	1415 ± 572	74	121	1022 ± 446	47	5	1120 ± 926	40
15–18 yrs	28	1566 ± 481	57	109	1474 ± 473	45	8	1588 ± 845	25
All ages	**108**	**1226 ± 546**	**58**	**344**	**1064 ± 526**	**45**	**20**	**1121 ± 784**	**35**
	Second Visit, Proposed Dose Followed								
	MMF + CSA			**MMF + TAC**			**MMF + another ISD or none**		
	No. of patients (n)	**Mean daily dose of MMF ± SD**	**Patients with 30 < AUC < 60, in %**	**No. of patients (n)**	**Mean daily dose of MMF ± SD**	**Patients with 30 < AUC < 60, in %**	**No. of patients (n)**	**Mean daily dose of MMF ± SD**	**Patients with 30 < AUC < 60, in %**
<5 yrs	12	887 ± 402	83	19	615 ± 331	63	1	800 ± 0	100
5–10 yrs	17	1001 ± 460	41	51	773 ± 315	69	4	670 ± 363	50
10–15 yrs	29	1521 ± 576	52	75	1042 ± 347	49	3	750 ± 177	33
15–18 yrs	20	1525 ± 510	65	78	1429 ± 515	64	5	1500 ± 612	60
All ages	**78**	**1311 ± 574**	**58**	**223**	**1080 ± 496**	**60**	**13**	**1018 ± 572**	**54**
	Second Visit, Proposed Dose Not Followed								
	MMF + CSA			**MMF + TAC**			**MMF + another ISD or none**		
	No. of patients (n)	**Mean daily dose of MMF ± SD**	**Patients with 30 < AUC < 60, in %**	**No. of patients (n)**	**Mean daily dose of MMF ± SD**	**Patients with 30 < AUC < 60, in %**	**No. of patients (n)**	**Mean daily dose of MMF ± SD**	**Patients with 30 < AUC < 60, in %**
<5 yrs	8	778 ± 404	25	12	519 ± 209	33	1	500 ± 0	0
5–10 yrs	8	934 ± 373	38	32	716 ± 464	41	1	500 ± 0	100
10–15 yrs	6	1333 ± 753	33	46	952 ± 437	48	2	1500 ± 707	50
15–18 yrs	8	1906 ± 626	63	31	1448 ± 494	52	3	1523 ± 1361	33
All ages	**30**	**1231 ± 689**	**40**	**121**	**974 ± 537**	**45**	**7**	**1224 ± 972**	**43**
Followed vs. not followed: *P* value			0.08			**0.006**			0.5

Bold indicates statistical difference at $p < 0.05$.
P values were calculated using the Fisher exact test.
ISD, immunosuppressant drug.

The larger proportion of patients treated with MMF and CSA and underexposed to MPA, observed at the first request (Fig. 1), may be due to the interaction of CSA with the enterohepatic circulation of MPA.[13,14] Biliary excretion of mycophenolic acid glucuronide from the hepatocytes involves the multidrug resistance-associated protein-2, which is inhibited by CSA.[15] Therefore, biliary excretion of mycophenolic acid glucuronide is impaired; consequently, a reduced amount of mycophenolic acid glucuronide is deglucuronidated by intestinal flora, less amount of MPA is reabsorbed, and MPA plasma concentrations are reduced.

In almost all age groups, the mean proposed dose, based on the first AUC, was significantly higher than the mean initial dose. Furthermore, the mean proposed doses were significantly different, depending on the concomitant immunosuppressants, and logically higher for patients on CSA.

The proportion of patients within the 30–60 mg h/L range at the second request when the proposed dose had been followed was not always higher, which was unexpected. However, smaller patient numbers may be the cause, for example, patients on CSA, with age groups <5, 5–10, 10–15, and 15–18 years old, comprised approximately 20 patients in the followed-dose subgroup and <10 patients in the other group. Notably, the patient group aged 10–15 years neither exhibited any improvement nor worsening of target attainment, irrespective of the drug combination. This might be due to poor compliance, often frequent in this age group.[16–20] Furthermore, the category with the higher percentage within the target range at the first request (CSA group, 58%), i.e., with the least room for improvement, demonstrated a smaller increase in target attainment at the subsequent visit. The percentage of AUCs within the target range at the second visit were significantly higher for all combinations of immunosuppressants when the proposed dose was followed.

Figure 2 is notable in several aspects; it reveals that during the first month, several patients are underexposed, and when the recommended dose is followed, MPA exposure

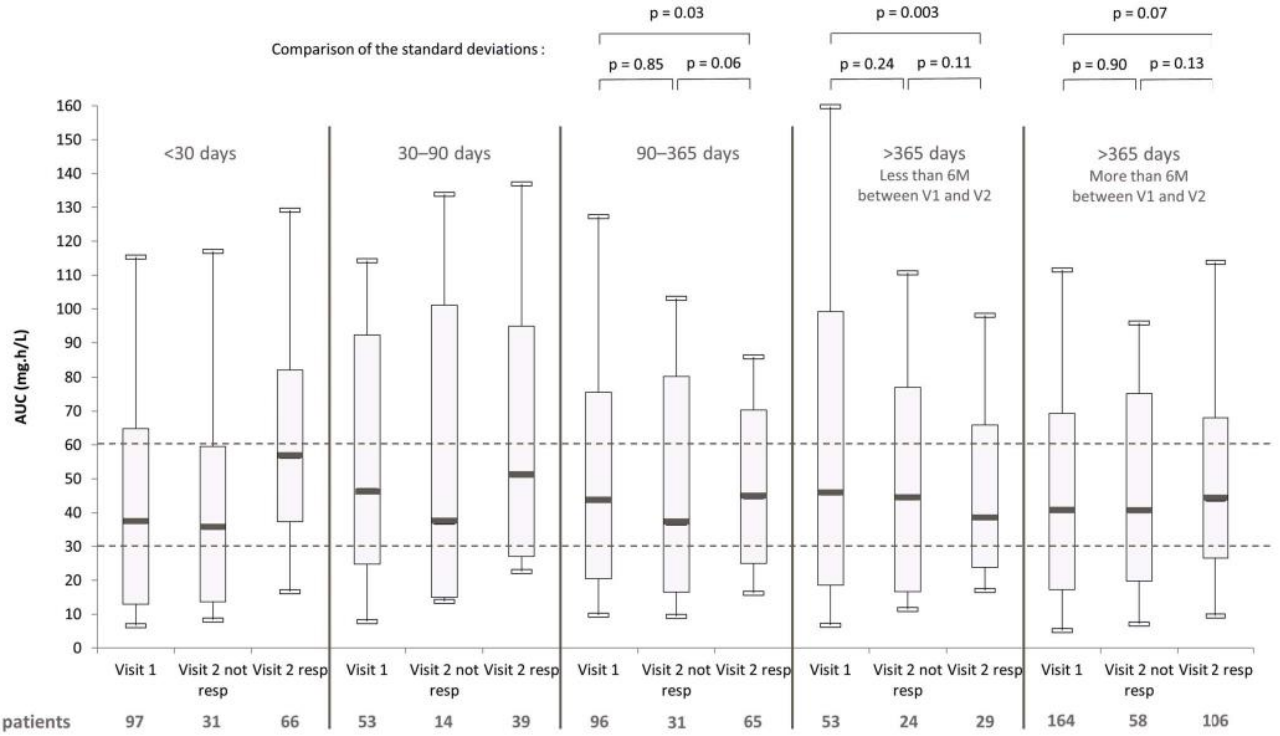


FIGURE 2. Distribution of MPA $AUC_{0-12\ h}$ in patients who benefited from consecutive AUC measurements (n = 463). Box limits indicate the first and third quartiles, and the central line is the median value. Lines extending from each box capture the range of the remaining data. Standard deviations were compared using the Fisher exact test. Visit 1 (V1), first AUC measurement; Visit 2 (V2) *not responding*, dose recommended at visit 1 *not followed* at visit 2; Visit 2 (V2) *responding*, dose recommended at visit 1 *followed* at visit 2.

is increased at the next visit. After 3 months, it reveals that the range of MPA is considerably large and narrows down at the first and second visits, respectively, with more AUCs in the therapeutic range, particularly when the recommended doses are followed and the first and second visits are closer in time. By contrast, when visits are >6 months apart (Fig. 2 right side), the frequency of underexposure is decreased between the first and second visits, although not that of overexposure. This may be explained by the increasing intraindividual variability in the long term. Moreover, the nadir of Cl/F at 4.5 months posttransplantation and Cl/F increase between the ages of 5 and 18 suggest that MPA AUC should be more frequently checked during these periods and MMF dose adjusted to compensate for these natural evolutions.

Another study on dose adjustment requests of TAC in pediatric KTRs revealed similar results; when the ISBA recommended dose was followed, the AUC distribution was narrower and a significantly higher proportion was within the target range.[21] This study was the first to present an overview of MPA exposure in pediatric KTRs on a large scale, especially in patients on MMF and TAC. In a study by Weber et al,[22] MPA AUCs were calculated using 3 concentrations (C_0, $C_{1\ h\ 15\ min}$, and $C_{4\ h}$) and multiple regression analysis. MMF with CSA was administered to 54 children twice a day at a total fixed dose of 1200 mg/m^2. The tendencies were similar to those observed in Figure 1 in the present study, with a high proportion of underexposed patients in the first month and several overexposed patients at ≥3 months, similar to adult patients. A smaller study (n = 9) reported overexposure to MPA in young KTRs at a later posttransplantation period (median = 55 months).[23] In 2003, David-Neto et al[24] studied 20 stable pediatric KTRs (aged ≤15 years) on MMF (80% with CSA) and estimated MPA exposure using 7 measures of plasma concentration (C_0, $C_{1\ h}$, $C_{2\ h}$, $C_{3\ h}$, $C_{4\ h}$, $C_{6\ h}$, and $C_{8\ h}$). The number of underexposed patients (15%) was similar to those in the present study; however, they reported less adequately exposed (30%) and more overexposed (55%) patients as well. Patients were given lower MMF doses (785 mg/m^2 per day for patients with approximately 1 m^2 of body surface area) than in the present study and they were studied at a later posttransplant period (25 ± 17 months). These studies confirmed the importance of using MMF dose adjustment instead of a fixed dose or a progressively decreasing dose scheme, whether at earlier or later posttransplant periods. When MMF was combined with CSA, most pediatric KTRs were underexposed to MPA, whereas when it was combined with TAC, the percentages of the underexposed and overexposed KTRs at the first visit were higher and more balanced (Fig. 1).

One limitation of this study was the absence of body weight data in the data set at each request because it was not necessary for the predictors; thus, calculating the mean doses in mg/kg and comparing the obtained results with others or providing dosing recommendations to maximize the probability of body weight–based AUC target attainment were not possible.

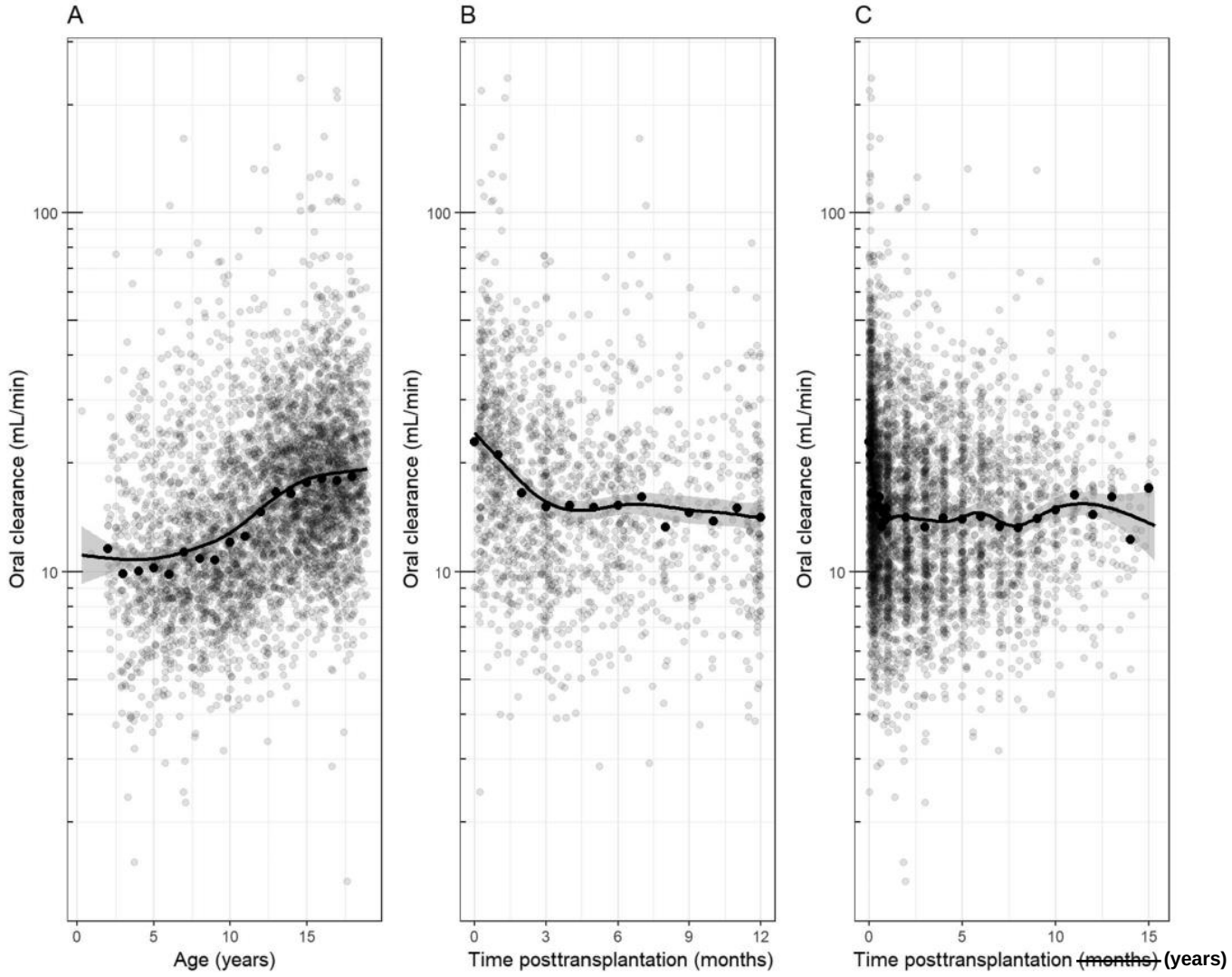


FIGURE 3. Oral clearance (dose/AUC) according to (A) Age, (B) time posttransplantation in months during the first year, and (C) time posttransplantation in years for all patients. Oral clearance is presented on a logarithmic scale. The curve represents the trend line from locally estimated scatterplot smoothing. The gray zone represents the 95% confidence interval. The black dots represent local medians.

CONCLUSION

This study revealed that before MMF dose adjustment based on MPA $AUC_{0-12\ h}$, large proportions of pediatric KTRs are not adequately exposed to MPA, irrespective of the age group and posttransplant period. Furthermore, adequate dose adjustment significantly increased the proportions of patients in the target range.

ACKNOWLEDGMENTS

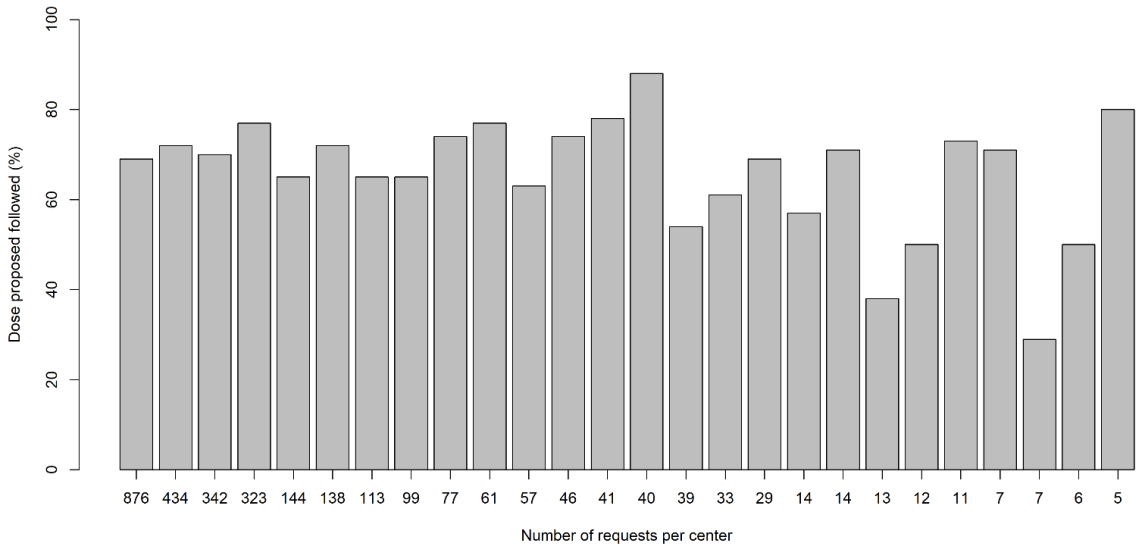
The authors thank the patients whose data were used in this study and their parents, and the physicians and clinical pharmacologists of the 51 hospitals worldwide who use the ISBA, for their trust. The authors thank Ms. Karen Poole for editing the manuscript.

REFERENCES

1. Woillard JB, Saint-Marcoux F, Monchaud C, et al . Mycophenolic mofetil optimized pharmacokinetic modelling, and exposure-effect associations in adult heart transplant recipients. *Pharmacol Res.* 2015;99:308–315.
2. Bergan S, Brunet M, Hesselink DA, et al. Personalized therapy for mycophenolate: consensus report by the international association of therapeutic drug monitoring and clinical toxicology. *Ther Drug Monit.* 2021; 43:150–200.
3. Tett SE, Saint-Marcoux F, Staatz CE, et al. Mycophenolate, clinical pharmacokinetics, formulations, and methods for assessing drug exposure. *Transplant Rev.* 2011;25:47–57.
4. van Gelder T, Hilbrands LB, Vanrenterghem Y, et al. A randomized double-blind, multicenter plasma concentration controlled study of the safety and efficacy of oral mycophenolate mofetil for the prevention of acute rejection after kidney transplantation. *Transplantation.* 1999;68:261–266.
5. Knight SR, Morris PJ. Does the evidence support the use of mycophenolate mofetil therapeutic drug monitoring in clinical practice? A systematic review. *Transplantation.* 2008;85:1675–1685.

6. Sommerer C, Müller-Krebs S, Schaier M, et al. Pharmacokinetic and pharmacodynamic analysis of enteric-coated mycophenolate sodium: limited sampling strategies and clinical outcome in renal transplant patients. *Br J Clin Pharmacol.* 2010;69:346–357.
7. Kuypers DRJ, de Jonge H, Naesens M, et al. Current target ranges of mycophenolic acid exposure and drug-related adverse events: a 5-year, open-label, prospective, clinical follow-up study in renal allograft recipients. *Clin Ther.* 2008;30:673–683.
8. Prémaud A, Debord J, Rousseau A, et al. A double absorption-phase model adequately describes mycophenolic acid plasma profiles in de novo renal transplant recipients given oral mycophenolate mofetil. *Clin Pharmacokinet.* 2005;44:837–847.
9. Prémaud A, Weber LT, Tönshoff B, et al. Population pharmacokinetics of mycophenolic acid in pediatric renal transplant patients using parametric and nonparametric approaches. *Pharmacol Res.* 2011;63:216–224.
10. Prémaud A, Meur YL, Debord J, et al. Maximum a posteriori bayesian estimation of mycophenolic acid pharmacokinetics in renal transplant recipients at different postgrafting periods. *Ther Drug Monit.* 2005;27:354–361.
11. Saint-Marcoux F, Woillard JB, Jurado C, Marquet P. Lessons from routine dose adjustment of tacrolimus in renal transplant patients based on global exposure. *Ther Drug Monit.* 2013;35:322–327.
12. Saint-Marcoux F, Vandierdonck S, Prémaud A, et al . Large scale analysis of routine dose adjustments of mycophenolate mofetil based on global exposure in renal transplant patients. *Ther Drug Monit.* 2011;33:285–294.
13. Trkulja V, Lalić Z, Nađ-Škegro S, et al. Effect of cyclosporine on steady-state pharmacokinetics of MPA in renal transplant recipients is not affected by the MPA formulation: analysis based on therapeutic drug monitoring data. *Ther Drug Monit.* 2014;36:456–464.
14. Filler G, Lepage N, Delisle B, Mai I. Effect of cyclosporine on mycophenolic acid area under the concentration-time curve in pediatric kidney transplant recipients. *Ther Drug Monit.* 2001;23:514–519.
15. van Gelder T. How cyclosporine reduces mycophenolic acid exposure by 40% while other calcineurin inhibitors do not. *Kidney Int.* 2021;100:1185–1189.
16. Obi Y, Ichimaru N, Kato T, et al. A single daily dose enhances the adherence to immunosuppressive treatment in kidney transplant recipients: a cross-sectional study. *Clin Exp Nephrol.* 2013;17:310–315.
17. Kaboré R, Couchoud C, Macher MA, et al. Age-dependent risk of graft failure in young kidney transplant recipients. *Transplantation.* 2017;101:1327–1335.
18. Foster BJ, Dahhou M, Zhang X, et al . Association between age and graft failure rates in young kidney transplant recipients. *Transplantation.* 2011;92:1237–1243.
19. Nevins TE. Non-compliance and its management in teenagers. *Pediatr Transplant.* 2002;6:475–479.
20. Dobbels F, Ruppar T, De Geest S, et al . Adherence to the immunosuppressive regimen in pediatric kidney transplant recipients: a systematic review. *Pediatr Transplant.* 2009;14:603–613.
21. Marquet P, Cros F, Micallef L, et al. Tacrolimus bayesian dose adjustment in pediatric renal transplant recipients. *Ther Drug Monit.* 2021;43:472–480.
22. Weber LT, Shipkova M, Armstrong VW, et al. The pharmacokinetic-pharmacodynamic relationship for total and free mycophenolic acid in pediatric renal transplant recipients: a report of the German study group on mycophenolate mofetil therapy. *J Am Soc Nephrol.* 2002;13:759–768.
23. Jacqz-Aigrain E, Khan Shaghaghi E, Baudouin V, et al. Pharmacokinetics and tolerance of mycophenolate mofetil in renal transplant children. *Pediatr Nephrol.* 2000;14:95–99.
24. David-Neto E, Araujo LMP, Sumita NM, et al. Mycophenolic acid pharmacokinetics in stable pediatric renal transplantation. *Pediatr Nephrol.* 2003;18:266–272.

Supplemental Digital Content



Supplemental Figure 1. Proportions of proposed doses followed between various centers (centers with only one or two requests are not represented)

II. Optimiser les estimations d'AUC en utilisant le Machine Learning et les simulations

II.1. Objectifs

Bien que l'aire sous la courbe des concentrations soit théoriquement le meilleur indice d'exposition pour l'évérolimus, il est peu utilisé. En effet, il nécessite de réaliser de nombreux prélèvements à intervalles réguliers entre deux prises du médicament, pour permettre de calculer l'AUC par méthode des trapèzes. De plus, peu de modèles PopPK existent, et encore moins de stratégies de prélèvement limité. Pour toutes ces raisons, cet indice d'exposition a été jusque-là peu étudié. Sur le site internet ABIS, un estimateur bayésien permet d'estimer ces AUC à partir de 3 prélèvements seulement, effectués juste avant la prise (C0), 1h et 2h après.

Les objectifs de cette étude étaient de : (i) améliorer les estimations d'AUC déjà fournies par l'estimateur bayésien MAP de Limoges, en entraînant un algorithme de ML sur les AUC estimées d'évérolimus et les trois concentrations sanguines mesurées, issues de profils de patients réels (données disponibles sur ABIS) ; (ii) utiliser un modèle PopPK issu de la littérature, autre que celui utilisé sur ABIS, pour créer des profils simulés et ajouter de la diversité dans les données d'apprentissage ; (iii) explorer les limites de l'utilisation de simulations dans l'entraînement d'un modèle de ML.

L'algorithme de ML XGBoost a été entraîné dans un premier temps sur les 508 AUC de patients réels, estimés sur ABIS. Puis, l'apprentissage a été enrichi avec 500 à 10 000 simulations issus d'un modèle PopPK de la littérature [131]. Les performances finales de chacun des modèles ont été évaluées à l'aide d'un jeu de données indépendant de 114 profils pharmacocinétiques complets (AUC de référence calculée par méthode des trapèze). Les indices de performance utilisés sont entre autres l'écart quadratique moyen (RMSE)[8] et le coefficient de détermination R^2.

II.2. Discussion

Dans cette étude, nous avons construit un modèle d'estimation de l'aire sous la courbe des concentrations d'évérolimus à partir de 3 prélèvements, en utilisant XGBoost, un algorithme de ML. Les prélèvements étaient réalisés à approximativement T0, T1h et T2h.

Dans un premier temps, nous avons utilisé exclusivement les profils de patients réels, puis nous avons fait des tests en ajoutant des jeux de simulations de tailles différentes (Article 2).

[8] RMSE : racine carrée de la moyenne des carrés des erreurs :

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(Predit_i - Réel_i)^2}{n}}$$

Tableau 1 : Résumé des stratégies d'apprentissage de l'Article 2 [132]

Les modèles *équivalents* en performances sont présentés à chaque fois dans une *même ligne*. Les performances présentées ici sont issues du jeu de données indépendant de 114 profils complets de concentrations (validation externe).

Base de données utilisée	RMSE (µg.h/L)	Base de données utilisée	RMSE (µg.h/L)	Base de données utilisée	RMSE (µg.h/L)
508 réels	18,0	1003 simulés	18,6		
508 réels + 500 simulés	12,5	2508 simulés	11,4	2508 simulés + 508 réels	12,2
		5016 simulés	10,8		

Avec notre meilleur modèle, l'erreur RMSE était très faible à 10,8 µg.h/L, par rapport aux valeurs à estimer : l'$AUC_{0\text{-}12h}$ moyenne du jeu de données de validation externe était de 96 µg.h/L (Tableau 1 de l'article).

Dans un deuxième temps, pour améliorer l'apprentissage, nous avons augmenté le nombre de profils simulés de 250 jusqu'à 15 000. Dans le jeu de test, issu des données simulées, l'erreur a diminué d'environ 30 µg.h/L, jusqu'à environ 6 µg.h/L (Figure 1 de l'article). Au-delà, les temps de calcul étaient importants,[9] particulièrement pour la validation croisée à 10 folds (voir la partie Séparation des jeux de données).

Nous avons ainsi volontairement testé les limites d'une telle approche en augmentant le nombre de simulations outre mesure. Notre objectif était de voir le seuil à partir duquel l'imprécision RMSE ne diminuerait plus sur un jeu indépendant. Avec le modèle de PopPK (issu de la littérature) utilisé pour les simulations, et le jeu de données externe issu de l'essai Everold, le nombre optimal de simulations était d'environ 5000 (Figure 1 de l'article). Pour 10 000 ou 15 000 simulations, l'erreur continuait de diminuer quand le modèle était évalué sur une partie des données simulées (jeu de test), mais elle augmentait progressivement dans le jeu de validation externe. Bien sûr, il est possible que ce seuil de 5 000 ne soit pas le même dans un contexte différent. Nous n'avons pas encore pu le comparer à d'autres valeurs de la littérature, car à ce jour peu de modèles de ML (pour l'estimation d'AUC) ont été entraînés sur des jeux de simulations croissants. En effet, cela nécessite à la fois des compétences en pharmacocinétique de population (pour l'utilisation du package mrgsolve sur le logiciel R), mais également en ML pour l'utilisation d'algorithmes comme XGBoost (optimisation des hyperparamètres, entraînement, sélection des variables etc.).

La pharmacocinétique de cet inhibiteur de mTOR a été moins étudiée que celle d'autres immunosuppresseurs. En conséquence, il existe moins de données et de modèles PK, limitant le développement d'outils d'estimation de l'AUC avec une stratégie d'échantillonnage limitée.

[9] Avec la configuration suivante : système d'exploitation Microsoft Windows 11 Famille ; processeur Intel® Core™ i7-8550U CPU @ 1.80GHz, 1992 MHz, 4 cœurs, 8 processeurs logiques ; mémoire physique (RAM) installée 8,00 Go.

Moes et al. ont construit un modèle utilisant deux points de concentrations C0 et C2h [131]. La performance de leur modèle était un peu moins bonne que notre meilleur modèle publié ici qui, lui, a été validé sur un jeu de données indépendant : R^2 = 0,90 et 0,96 respectivement. Robertsen et al. ont aussi développé un outil d'estimation de l'AUC de l'évérolimus à partir d'un nombre limité de prélèvements de sang total [22]. La performance de leur modèle était meilleure que celle présentée ici, mais il a été validé sur un très petit nombre de profils : RMSE normalisé[10] = 9,9 % (n = 4), versus 10,3 % (n = 114 d'un jeu de données indépendant). Enfin, dans le travail de Zwart et al. [23], les erreurs étaient plus élevées, et les AUC ayant servi de référence étaient calculées à partir de moins de points de concentrations : n = min-max 4-7, versus n = 10-12 dans notre étude.

En résumé, nous avons perfectionné l'apprentissage d'un algorithme de ML pour estimer l'AUC de l'évérolimus à partir d'une stratégie de prélèvement limitée. Deux études similaires ont été précédemment publiées par notre équipe de recherche Inserm 1248, pour le tacrolimus. Dans l'une, à la différence de l'évérolimus, de très nombreux profils réels étaient disponibles pour l'apprentissage (n = 6449) [133]. Ils avaient permis d'entraîner des modèles utilisant seulement 2 concentrations et plusieurs covariables. Avec 2 prélèvements, les performances sur les jeux de données indépendants étaient équivalentes ou meilleures avec XGBoost qu'avec un estimateur MAP. Dans la deuxième étude sur le tacrolimus, un nombre très élevé de simulations (n = 4192) avaient été utilisées [134]. Les résultats montraient des RMSE normalisées à 4,6 % (2 prélèvements) et 2,6 % (3 prélèvements) dans les bases de données tests (profils simulés à partir du même modèle PopPK que pour l'apprentissage). Alors que dans les jeux de validation externe, l'imprécision était un peu plus élevée : RMSE normalisée de 8,1 à 11,5 %. Cette différence de performance entre la base de données de test et les bases de données externes était probablement la conséquence d'un surapprentissage, même si cela n'est pas évoqué dans cet article. Des jeux de données simulés de taille différente n'ont pas été testées pour l'investiguer. Toutefois, l'imprécision mise en avant dans le résumé principal de ce deuxième article sur le tacrolimus était (en toute transparence) celle des jeux de données indépendants.

Dans notre cas, nous sommes partis d'un nombre *limité* de profils de patients réels et nous avons utilisé un nombre *croissant* de simulations. Nous avons sciemment poussé ce raisonnement jusqu'à ses limites, pour trouver le nombre approximatif de simulations qui entraînerait du surapprentissage.

Cet outil de calcul de l'AUC permettra d'améliorer le suivi de l'exposition à l'évérolimus, et de *mieux étudier* son influence sur des critères de jugements majeurs, comme le rejet.

[10] $RMSE\ normalisé = \frac{RMSE}{Moyenne\ des\ AUC\ de\ référence}$ (critère de performance comparable entre différents médicaments, où les AUC observées peuvent atteindre des moyennes différentes)

II.3. Article 2

Machine learning algorithms to estimate everolimus exposure trained on simulated and patient pharmacokinetic profiles

CPT Pharmacometrics Syst Pharmacol. 2022

Received: 11 February 2022 | Revised: 26 April 2022 | Accepted: 28 April 2022

DOI: 10.1002/psp4.12810

ARTICLE

Machine learning algorithms to estimate everolimus exposure trained on simulated and patient pharmacokinetic profiles

Marc Labriffe[1,2] | **Jean-Baptiste Woillard**[1,2] | **Jean Debord**[1,2] | **Pierre Marquet**[1,2]

[1]Pharmacology & Transplantation, INSERM U1248, Université de Limoges, Limoges, France

[2]Department of Pharmacology, Toxicology and Pharmacovigilance, CHU de Limoges, Limoges, France

Correspondence
Pierre Marquet, Department of Pharmacology, Toxicology and Pharmacovigilance, University Hospital of Limoges, CBRS, 2 rue Bernard Descottes, 87000 Limoges, France.
Email: pierre.marquet@unilim.fr

Funding information
No funding was received for this work.

Abstract

Everolimus is an immunosuppressant with a small therapeutic index and large between-patient variability. The area under the concentration versus time curve (AUC) is the best marker of exposure but measuring it requires collecting many blood samples. The objective of this study was to train machine learning (ML) algorithms using pharmacokinetic (PK) profiles from kidney transplant recipients, simulated profiles, or both types, and compare their performance for everolimus $AUC_{0\text{-}12h}$ estimation using a limited number of predictors, as compared to an independent set of full PK profiles from patients, as well as to the corresponding maximum a posteriori Bayesian estimates (MAP-BE). XGBoost was first trained on 508 patient interdose AUCs estimated using MAP-BE, and then on 500–10,000 rich interdose PK profiles simulated using previously published population PK parameters. The predictors used were: predose, ~1 h, and ~2 h whole blood concentrations, differences between these concentrations, relative deviations from theoretical sampling times, morning dose, patient age, and time elapsed since transplantation. The best results were obtained with XGBoost trained on 5016 simulated profiles. AUC estimation achieved in an external dataset of 114 full-PK profiles was excellent (root mean squared error [RMSE] = 10.8 μg*h/L) and slightly better than MAP-BE (RMSE = 11.9 μg*h/L). Using more profiles (n = 10,035) did not improve the ML algorithm performance. The contribution of mixing patient and simulated profiles was significant only when they were in balanced numbers, with ~500 for each (RMSE = 12.5 μg*h/L), compared with patient data alone (RMSE = 18.0 μg*h/L).

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Assessing everolimus area under the concentration-time curve (AUC) requires either collecting many blood samples or using a pharmacokinetic (PK) model and Bayesian estimator with a few blood samples. Machine learning (ML) algorithms represent an alternative, provided they can be trained on large enough databases.

WHAT QUESTION DID THIS STUDY ADDRESS?
It evaluated the contribution and limits of simulated data to train ML models to estimate everolimus $AUC_{0\text{-}12h}$.
WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?
An optimal amount of simulated data ($n = 5016$ PK profiles) optimized XGBoost $AUC_{0\text{-}12h}$ prediction, even rendering patient data useless, and yielded better performance than Bayesian estimation.
HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?
When limited data is available to train ML algorithms, simulations can be used. However, too many simulated data expose to overfitting, highlighting the need for independent patient datasets for external validation.

INTRODUCTION

Everolimus is an inhibitor of the mammalian target of rapamycin (mTOR) activity, in particular in lymphocytes.[1] It is a non-nephrotoxic drug that shows a synergistic immunosuppressive effect with calcineurin inhibitors (CNIs).[2,3] It is characterized by a narrow therapeutic range and a large interindividual variability requiring concentration-based dose adjustments, similar to CNIs. Therapeutic drug monitoring (TDM) is therefore recommended—or, in certain countries, compulsory—for everolimus and, due to its high distribution in erythrocytes, dose individualization is generally based on trough whole blood concentrations.[4]

Two main markers are currently available to individually adjust everolimus dose: the trough blood level (C0), which is widely used for practical and economic reasons, although it has inconsistently been associated with clinical outcomes, and the interdose area under the curve ($AUC_{0\text{-}12h}$),[5] which reflects overall exposure and is theoretically a better predictor of the drug pharmacodynamics. Kovarik et al.[6] actually found exposure-response relationships between everolimus AUC and the incidence of thrombocytopenia, hypertriglyceridemia, and hypercholesterolemia. In the same study, he measured the interindividual (coefficient of variation [CV] = 85.4%) and the intra-individual (interoccasion) (CV = 40.8%) variability of the AUC, suggesting that TDM is needed and feasible, respectively. Another study showed that the interindividual variability is larger for everolimus C0 than $AUC_{0\text{-}12h}$ (CV% = 55% and 31%, respectively), as is the intra-individual variability (45% and 27%, respectively).[7] However, the interdose AUC is more difficult to measure than C0 because it requires collecting and analyzing many samples. In practice, everolimus AUC has been estimated using population pharmacokinetic (PopPK) models and maximum a posteriori Bayesian estimation (MAP-BE) based on limited sampling strategies.[8] In 2005, the Immunosuppressant Bayesian Dose Adjustment (ISBA) expert system and website (https://abis.chu-limoges.fr/login) were launched to share tools able to estimate the interdose AUC of immunosuppressants using MAP-BE on the basis of three blood samples and some patient characteristics (type of graft, age, post-transplantation period, and drug measurement assay).[9] In 2018, a new everolimus model was made available on ISBA, where each request posted is validated in less than 48 h by a trained pharmacologist.

Over the last 2 decades, machine learning (ML) has been successfully used in many applications in pharmacology, thanks to the huge and ever-increasing amount of data and computational power as well as to the improvement of learning algorithms.[9,10] Extreme gradient boosting (XGBoost) is an ML algorithm where simple regression trees are iteratively built by finding split values among all input variables to minimize prediction error. The iterative process constructs an additional regression tree of the same structure to minimize the residual errors of the previous regression tree.[11] We found that XGBoost was particularly suited to estimate the AUC of other immunosuppressive drugs using limited sampling strategies and covariates.[12,13] For tacrolimus in particular, we even trained in parallel such algorithms on massive simulated data rather than actual patient data, showing again better performance than usual MAP-BE.[14] This was an important finding because, for many drugs such as everolimus, there is not enough patient data available to train ML algorithms. However, the full potential of simulated data combined with patient data has not been explored yet. Is there an optimal number of simulations? Is patient data, even in rather low volume, still useful if a potentially infinite number of PK profiles can be simulated? If so, is there an optimal balance between patient and simulated data?

The objective of this study was to compare different combinations of patient and simulated PK profiles for the training of an XGBoost algorithm able to estimate everolimus $AUC_{0\text{-}12h}$ using a limited number of predictors. The true performance was evaluated in external validation datasets of full concentrations profiles from kidney transplant recipients, and then compared to that of MAP-BE in the same datasets.

METHODS

Patients and actual data

The everolimus AUC estimation and dose recommendation requests received on our ISBA website since 2018 for recipients of a renal transplant were extracted and cleaned using the Tidyverse framework. Data collection was approved by the regional ethics committee, and all patients gave their informed consent to participate in the study (EudraCT number 2006–0068 32–23 and 2009–0135 41–28). Blood was collected at three sampling times at least: predose (C0), ~60 min (30–100 min, C1), and ~120 min (115–220 min, C2) after drug intake. Everolimus blood levels were measured using high-performance liquid chromatography coupled to tandem mass spectrometry. The other predictors available were the morning dose of everolimus, the time elapsed between transplantation and everolimus blood sampling, and patient age. The code used for data cleaning and data that support the findings of this study are available upon request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Simulated data

We used the parameters of a previously published pharmacokinetic (PK) model developed for everolimus in a population of adult kidney transplant recipients.[16] PK profiles were simulated at steady-state over a 12-h interval, uniformly for different drug doses (0.5, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.25, 3.5, 4, and 4.5 mg), using the mrgsolve R package.[17] The proportional error was significantly diminished to 0.01% (as compared to 13.9% in the original paper) to obtain less noisy, smoother simulated PK profiles, deliberately neglecting measurement errors to optimize algorithm training on unaltered, "true" AUC values. However, at the prediction step, gaussian noise with mean = 1.2% and SD = 3,9% (minimum = 0.0% and maximum = 130.9%) was randomly added to the simulated C1 and C2 sampling times, using the sdcMicro R package.[18] The aim was to introduce uncertainty on input data so as to observe the algorithm prediction performance in more realistic conditions. In addition, we kept the interindividual variability of the PK parameters described in the initial study (eta values), as well as that brought by the most important covariate, the ideal body weight. Indeed, in the original model, apparent volume of distribution of the central compartment after oral administration (V1/F) was a function of the ideal body weight. We simulated ideal body weight values using a truncated random normal distribution with mean ± SD (minimum-maximum) = 68 ± 7.5 (52–83) kg (in accordance with the original article).

Datasets and analysis strategy

The present study used supervised learning from different training datasets to predict the interdose AUC, whose reference values had been obtained either through our ISBA expert system using MAP-BE and three everolimus blood concentrations for kidney transplant patients; or using the trapezoidal rule with the PKNCA R package for the simulated profiles.[19] In order to maximize the diversity of the training sets when patient and simulated data were mixed, the simulated profiles were not obtained using our in-house PK model but using a model from the literature.[16] We trained many ML algorithms, based on patient, simulated, or mixed data. Each training dataset was used in turn to build an XGBoost algorithm, tune the hyperparameters, and evaluate its performance by a single 10-fold cross-validation (random partition of the training set into 10 parts). The algorithms were then evaluated on independent subsets of the training sets by calculating the root mean square error (RMSE; expressed in µg*h/L) between the estimated and reference AUCs. Finally, the different algorithms were comparatively evaluated using two independent datasets of everolimus full PK profiles in kidney transplant recipients.

Feature engineering

Everolimus blood concentrations (whether actually measured or simulated) were divided into three theoretical time classes: concentrations at trough (C0 sampled at t = 0 min), 1 h (C1 sampled between 30 and 100 min), and 2 or 3 h (C2 sampled between 115 and 220 min). New variables were drawn for times 1 and 2 h corresponding to the relative deviation with respect to the theoretical times. For instance, if the sampling time was 1.06 h, the relative time difference with the theoretical time 1 h was (1.06–1)/1 = 0.06. Other predictors corresponding to the differences between the concentrations C1–C0, C1–C2, and

C2–C0 were created to add information about potentially delayed absorption peaks. Finally, the features tested as predictors of the interdose AUCs in the training set from actual patients were: patient age, time elapsed between transplantation and everolimus blood sampling, everolimus morning dose, everolimus concentrations at times 0, 1 h, and 2 h, relative deviation from the theoretical times, and differences between concentrations. In the training set of simulated profiles, as well as in the mixed training sets, the potential predictors were limited to: everolimus concentrations at times 0, 1 h, and 2 h, relative deviation from the theoretical times, differences between concentrations, and everolimus morning dose.

Exploratory data analyses

A correlation matrix and scatterplots were drawn to explore the correlations between AUC and predictors in the actual patient dataset, using the GGally R package.[20]

Preprocessing of the data

For all the ML analyses, the tidymodels framework was used. No preprocessing was applied to the data because XGBoost methods do not require normalization prior to analysis. There were no missing data in the predictors. Data splitting between training datasets (75%) and test datasets (25%) was performed by random selection of patients (or simulated cases).

Training of XGBoost algorithms

The algorithms were tuned by searching the hyperparameter combination associated with the lowest RMSE and highest R^2 between estimated and reference AUC values, using 10-fold cross-validation. In brief, the best combination of hyperparameters was investigated in 90% of each training dataset in turn (analysis subset) and evaluated in the remaining 10% (assessment subset) and this process was repeated 10 times by circular permutation. The hyperparameters tuned among a grid of 30 random combinations were: the number of predictors randomly sampled at each split (mtry, between 1 and 11), the minimum number of data points required for the node to be split further (min_n between 1 and 40), the maximum depth of the tree (tree_depth, between 1 and 15), and the rate at which the boosting algorithm adapted from iteration-to-iteration (learn_rate, between 0 and 0.08). In a second time, the best hyperparameter combinations were evaluated by means of another set of 10-fold cross-validation to assess the mean RMSE and R^2 and their SDs in the corresponding training dataset and draw the scatter plots of estimated versus reference AUC. Finally, AUC estimation was evaluated in the respective test datasets by calculating RMSE, R,[2] normalized RMSE (RMSE divided by the mean of reference AUCs), relative mean prediction error (MPE), as well as through the number and proportion of estimates with absolute MPE greater than 20%. The code used for the simulation of PK profiles and XGBoost training is provided as Supplementary Text S1.

External evaluation of machine learning AUC estimates by comparison with full PK profiles, and comparison with maximum a posteriori Bayesian estimates

The independent validation dataset comprised full PK profiles from the PIGREC trial (NCT00812786; 0, 0.33, 0.66, 1, 1.5, 2, 3, 4, 6, 8, 9, and 12 h postdose) and from the Everold trial (NCT01028092; 0, 0.33, 0.75, 1, 1.66, 2, 4, 6, 8, 10, and 12 h postdose). Concentrations at 0, 1, and 2 h, everolimus dose, blood sampling times, and time elapsed between transplantation and everolimus blood sampling were extracted from the independent PK databases to predict the AUC using the ML algorithms as compared with the MAP-BE used in ISBA. The full concentration profiles were used to calculate the trapezoidal AUC (chosen as the reference) using the DescTools package.[21] The performance of the ML algorithms and of MAP-BE was evaluated by comparing the estimated AUCs to the trapezoidal AUCs in terms of RMSE and relative MPE, and the proportion of bias out of the ±20% interval. Additionally, the scatter plots of predicted versus reference AUCs and residuals versus predicted AUCs were drawn on the same graph for visual comparison of the different approaches.

RESULTS

Patients and data

The cleaned dataset extracted from ISBA used as patient data training set consisted of 508 everolimus $AUC_{0\text{-}12h}$ from 177 patients. The characteristics of the training and test sets of patient data are reported in Table 1. The median $AUC_{0\text{-}12h}$ was 101 (interquartile range [IQR] 73, 142] μg*h/L. The independent validation patient dataset comprised 114 PK profiles of 10–12 samples and in this group the coefficient of determination R^2 between C0 and the reference trapezoidal $AUC_{0\text{-}12h}$ ($n = 114$) was only 0.776.

TABLE 1 Characteristics of the features used for the training and validation of the first XGBoost algorithm based on 508 patient pharmacokinetic profiles

	Train set (n = 381)	Test set (n = 127)	External validation set (n = 114)
Time between transplantation and tacrolimus blood concentrations, months	3.95 [1.97, 11.84]	3.95 [1.97, 11.84]	14.76 [4.97, 105.90]
$AUC_{0\text{-}12h}$, µg*h/L	102 [74, 142]	101 [73, 145]	96 [69, 125]
Patient age, year	47 [35, 57]	47 [39, 57]	50 [40, 59]
Morning dose, mg	1.50 [0.75, 1.50]	1.50 [0.75, 1.50]	1.00 [0.56, 2.00]
Trough level (C0), µg/L	5.4 [3.7, 7.9]	5.5 [3.5, 8.6]	5.4 [3.9, 7.4]
Concentration at 1 h; C1, µg/L	14.2 [9.2, 20.5]	13.3 [8.7, 19.9]	15.2 [10.2, 21.0]
Concentration at 2 h; C2, µg/L	13.2 [9.0, 18.4]	12.9 [9.2, 18.3]	11.5 [8.4, 15.6]
Deviation from the 1-h theoretical time, %	0 [0, 0]	0 [0, 0]	0 [0, 0]
Deviation from the 2-h theoretical time, %	0 [0, 4]	0 [0, 2]	0 [0, 0]
Concentration difference between C1 and C0	8.5 [4.2, 13.1]	7.2 [3.3, 11.9]	9.5 [5.6, 14.1]
Concentration difference between C1 and C2	1.5 [−1.3, 4.6]	0.3 [−2.1, 3.6]	3.6 [0.6, 6.4]
Concentration difference between C2 and C0	7.1 [4.7, 10.8]	7.3 [4.3, 10.0]	5.8 [3.9, 9.1]
Reference AUCs: number of samples	3	3	10–12
Reference AUCs: method used	Same MAP-BE		Trapezoidal rule

Note: Medians [interquartile ranges] are presented here.

Abbreviations: AUC, area under the curve; ISBA, Immunosuppressant Bayesian Dose Adjustment; MAP-BE, maximum a posteriori Bayesian estimation; XGBoost, extreme gradient boosting, an optimized gradient boosting machine learning method.

Exploratory data analyses

The correlation matrix between everolimus $AUC_{0\text{-}12h}$ and predictors from patient PK profiles is presented in Figure S1, showing that the strongest correlations (>0.8) were between $AUC_{0\text{-}12h}$ and C0 or C2h.

XGBoost algorithms, training, and test sets

The best-tuned hyperparameter values for each algorithm are presented in Table S1. The results in the training sets obtained after 10-fold cross-validation and in the respective test sets are shown in Table 2. Among the test sets, the lowest RMSE (6.7 µg*h/L) was obtained using 10,035 simulated profiles.

External evaluation versus the trapezoidal AUC in an independent dataset

The best results (RMSE = 10.8 µg*h/L) were obtained using 5016 simulated profiles without patient data (Table 2). Focusing on the simulated profiles, Figure 1 presents the performances of XGBoost in the training and the external validation datasets according to the number of simulations used (including additional models trained on n = 250, 500 or 15,051 simulated profiles). It shows that the higher the number, the better the performance in the training set, whereas in the independent dataset RMSE followed a U shape curve with a minimal value for 5016 simulations.

Figure 2 presents the scatter plots and residual plots of estimated versus reference AUCs in the external validation dataset for four models: the algorithm based on the patient data only (n = 508); the best model mixing patient and simulated data (n = 508 and 500, respectively); the best model using only simulated data (n = 5016), and MAP-BE for comparison. There was no systematic bias. For our best model (5016 simulations), we also explored the possibility of adding more variability in the sampling times (Table S3), and it negatively affected the results in the test set, and a little bit in the validation set.

DISCUSSION

In this work, based on our previous experience with ML tools to estimate overall exposure to other immunosuppressive drugs,[13-15] we used XGBoost ML algorithms to estimate the interdose AUC of everolimus in renal transplant recipients. Because we had a more limited training dataset with actual patient data than with other immunosuppressants, we trained ML algorithms on patient, simulated, and mixed data and compared the estimates obtained in an independent database from kidney

TABLE 2 Performance of the XGBoost algorithms at estimating everolimus $AUC_{0\text{-}12h}$ in the different sorts of training, testing, and external validation datasets

		Train set (n = 75%)	Test set (n = 25%)	External validation set (n = 114 full PK profiles)	
		XGBoost	XGBoost	XGBoost (n = 114)	MAP-BE (n = 94[a])
508 patient PK profiles	RMSE, µg*h/L	15.2	15.4	18.0	11.9
	Normalized RMSE (%)	13.5	13.8	17.2	11.2
	R^2	0.921	0.922	0.873	0.952
	Relative MPE (%)	1.9	4.5	4.5	3.0
	Number of MPE out of the ±20% interval n	41 (10.8%)	20 (15.7%)	17 (14.9%)	7 (7.4%)
500 simulated + 508 patient PK profiles	RMSE, µg*h/L	32.1	23.3	12.5	11.9
	Normalized RMSE (%)	24.9	17.2	11.9	11.2
	R^2	0.880	0.942	0.939	0.952
	Relative MPE (%)	−0.4	0.8	0.0	3.0
	Number of MPE out of the ±20% interval n	50 (6.6%)	26 (10.3%)	5 (4.4%)	7 (7.4%)
1003 simulated PK profiles	RMSE, µg*h/L	18.5	19.0	18.6	11.9
	Normalized RMSE (%)	12.6	12.8	17.8	11.2
	R^2	0.970	0.970	0.919	0.952
	Relative MPE (%)	1.7	1.6	9.4	3.0
	Number of MPE out of the ±20% interval n	39 (5.2%)	13 (5.2%)	22 (19.3%)	7 (7.4%)
1003 simulated + 508 patient PK profiles	RMSE, µg*h/L	19.2	10.7	14.1	11.9
	Normalized RMSE (%)	13.6	7.8	13.4	11.2
	R^2	0.967	0.986	0.924	0.952
	Relative MPE (%)	0.5	0.0	1.2	3.0
	Number of MPE out of the ±20% interval n	50 (4.4%)	17 (4.5%)	8 (7.0%)	7 (7.4%)
2508 simulated PK profiles	RMSE, µg*h/L	13.4	15.1	11.4	11.9
	Normalized RMSE (%)	8.6	10.2	10.9	11.2
	R^2	0.987	0.982	0.951	0.952
	Relative MPE (%)	0.1	0.1	1.4	3.0
	Number of MPE out ofthe ±20% interval n	8 (0.4%)	4 (0.6%)	8 (7.0%)	7 (7.4%)
2508 simulated + 508 patient PK profiles	RMSE, µg*h/L	12.5	14.7	12.2	11.9
	Normalized RMSE (%)	8.5	10.2	11.7	11.2
	R^2	0.987	0.981	0.942	0.952
	Relative MPE (%)	0.3	0.2	2.2	3.0
	Number of MPE out of the ±20% interval n	39 (1.7%)	17 (2.3%)	7 (6.1%)	7 (7.4%)
5016 simulated PK profiles	RMSE, µg*h/L	14.1	11.2	10.8	11.9
	Normalized RMSE (%)	9.3	7.3	10.3	11.2
	R^2	0.985	0.990	0.956	0.952
	Relative MPE (%)	0.1	0.1	1.6	3.0
	Number of MPE out of the ±20% interval n	9 (0.2%)	2 (0.2%)	7 (6.1%)	7 (7.4%)

TABLE 2 (Continued)

		Train set (n = 75%)	Test set (n = 25%)	External validation set (n = 114 full PK profiles)	
		XGBoost	XGBoost	XGBoost (n = 114)	MAP-BE (n = 94[a])
5016 simulated + 508 patient PK profiles	RMSE, µg*h/L	11.1	9.2	12.7	11.9
	Normalized RMSE (%)	7.4	6.4	12.1	11.2
	R^2	0.990	0.992	0.939	0.952
	Relative MPE (%)	0.2	0.3	2.7	3.0
	Number of MPE out of the ±20% interval n	44 (1.1%)	16 (1.2%)	6 (5.3%)	7 (7.4%)
10,035 simulated PK profiles	RMSE, µg*h/L	7.6	6.7	12.6	11.9
	Normalized RMSE (%)	5.0	4.3	12.1	11.2
	R^2	0.996	0.997	0.942	0.952
	Relative MPE (%)	0.0	0.0	−1.2	3.0
	Number of MPE out of the ±20% interval n	3 (0.0%)	0 (0.0%)	7 (6.1%)	7 (7.4%)
10,035 simulated + 508 patient PK profiles	RMSE, µg*h/L	7.6	7.5	13.7	11.9
	Normalized RMSE (%)	5.0	4.9	13.1	11.2
	R^2	0.996	0.996	0.929	0.952
	Relative MPE (%)	0.1	0.1	2.6	3.0
	Number of MPE out of the ±20% interval n	41 (0.5%)	13 (0.5%)	9 (7.9%)	7 (7.4%)

Note: The performance of the MAP-BE actually used in the online ISBA expert system is displayed here, in the last column of the table, for comparison purposes.

Abbreviations: AUC, area under the curve; ISBA, Immunosuppressant Bayesian Dose Adjustment; MAP-BE, maximum a posteriori Bayesian estimation; MPE, mean prediction error; Normalized RMSE, root mean square error divided by the mean of reference AUCs; PK, pharmacokinetic; RMSE, root mean square error; XGBoost, extreme gradient boosting, an optimized gradient boosting machine learning method.

[a]For 20 profiles, MAP-BE could not be used because the morning dose was missing.

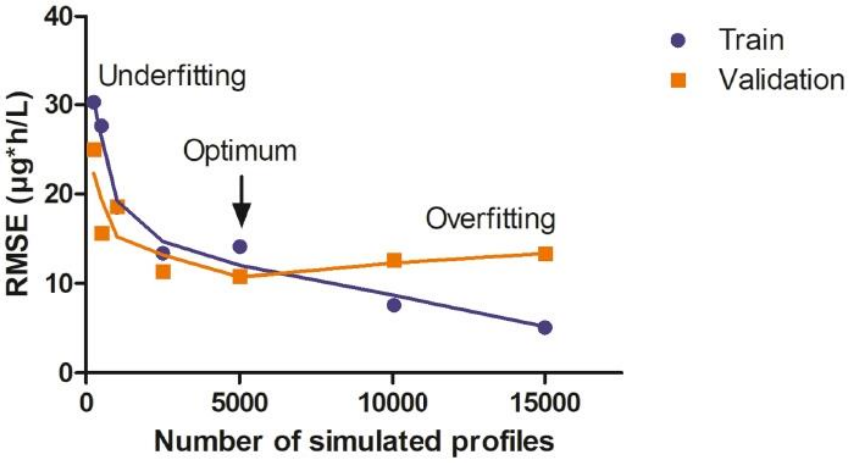


FIGURE 1 Plot of everolimus $AUC_{0\text{-}12h}$ prediction RMSE in the training (blue) and the external validation (orange) datasets, according to the number of simulations used to train the XGBoost algorithm. Points represent the performance of each XGBoost model, lines are a smoothed representation of trends. $AUC_{0\text{–}12h}$, 0–12-h area under the concentration-time curve; RMSE, root mean square error; XGBoost, extreme gradient boosting, an optimized gradient boosting machine learning method

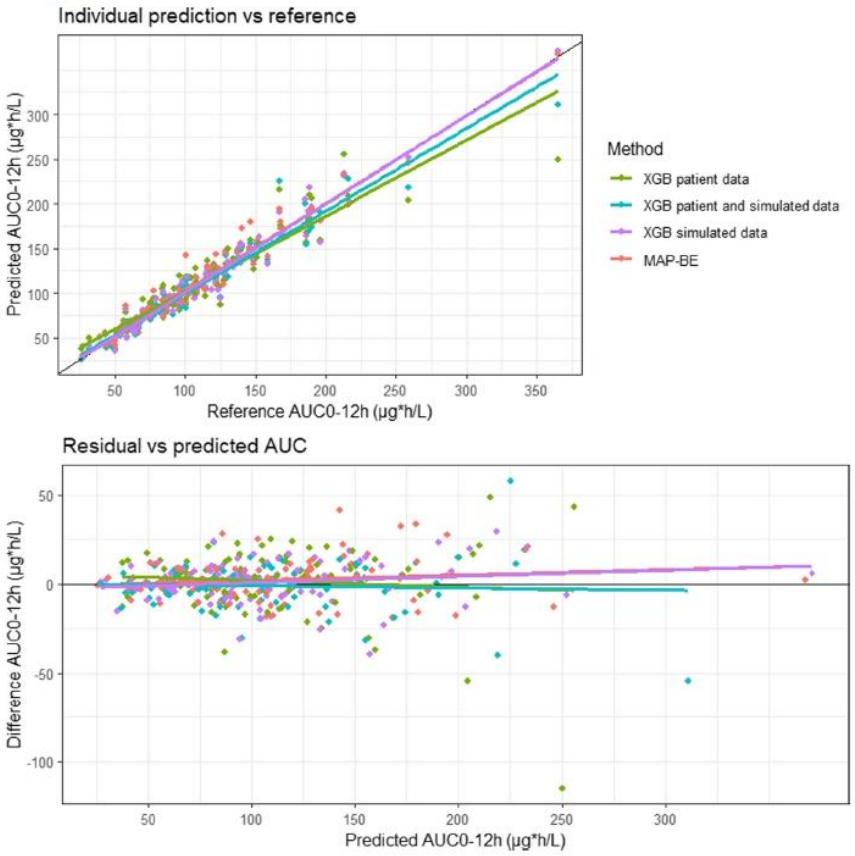


FIGURE 2 Scatter plots and residual plots of machine learning predicted versus reference everolimus $AUC_{0\text{-}12h}$ in the external validation dataset. The thin black line represents $y = x$. The colored lines were obtained by linear regression for each version of the XGBoost algorithm: in green, the model trained on patient data only ($n = 508$); in blue, the model trained on a balanced mix of patient ($n = 508$) and simulated ($n = 500$) data; in purple the best model, based on simulated data only ($n = 5016$); in red, the MAP-BE currently available through our online expert system ISBA. AUC, area under the curve; ISBA, Immunosuppressant Bayesian Dose Adjustment; MAP-BE, maximum a posteriori Bayesian estimation; XGBoost, extreme gradient boosting, an optimized gradient boosting machine learning method

transplant recipients with extensive-sampling with trapezoidal AUCs (reference AUCs) and MAP-BE AUC estimates based on a three-point limited sampling strategy, as used by our ISBA expert system. Different sizes of simulated training datasets were therefore compared based on several indicators, but primarily imprecision (i.e., RMSE). The performances of the ML algorithms trained on 5016 simulations without patient data yielded the best results. However, RMSE represents imprecision in the dataset and cannot be interpreted as the absolute error in a given patient. In our study, Figure 2 shows that absolute errors were lower for smaller reference AUCs.

The results in the training sets obtained after 10-fold cross-validation and in the respective test sets (Table 2) gradually improved with the number of simulated profiles (from 1003 up to the 10,035), yielding RMSE from 19.0 down to 6.7 µg*h/L in the test sets. This apparently unlimited decrease in RMSE was a sign of overfitting. When we evaluated the mixing of simulated and patient data, we noted that adding 500 simulated to the 508 patient profiles seemed to penalize the model as compared to patient data alone (RMSE = 23.3 µg*h/L and 15.4 in the test set, respectively). This is probably due to the wide diversity of profiles to be handled in this fairly small training set. In contrast, adding the 508 patients' profiles to 10,035 simulations increased the RMSE from 6.7 to 7.5 µg*h/L in the test set. This was a second sign of model overfitting due to the huge number of simulated profiles in the training set.

In a second step, all the models were externally evaluated using as references the trapezoidal AUCs of an independent patient dataset. Adding the 508 patient AUCs to the 2508, 5016, or 10,035 simulated profiles for training did not improve the performances of the algorithm at this validation step. With 5000 simulations or less, prediction RMSE was roughly equivalent in the training and the validation datasets. With 10,000 simulations or more, RMSE was still decreasing in the training datasets, whereas it was slowly increasing in the independent patient dataset, showing overfitting to the parametric model used for simulations.

In addition, as shown in Figure 2, the two algorithms trained on patient data (even partially) did worse for the highest AUC values than those trained only on simulated data ($n = 5016$), or than MAP-BE. The lowest RMSE (10.8 µg*h/L) in the external validation dataset (optimum) was obtained with 5016 simulated profiles. However, this precise optimal number of simulated profiles may not be generalized to all types of datasets, depending on the type of PK profiles, interindividual variability, data quality, etc.

It is worth noting that C0 values were not so well-correlated with the reference AUCs (R^2 = 0.776) calculated using the trapezoidal rule with greater than or equal to 10 samples from an unprecedented number of full PK profiles in kidney transplant recipients (114 profiles at median 14.8 [IQR = 5.0, 105.9] months post-transplantation in our independent validation database). Estimating AUC from C0 could therefore lead to great uncertainty (Table S2). Chan et al. compared C0 to incomplete AUC ($AUC_{0\text{-}5}$) in 92 patients at 1, 3, and 6 months post-transplantation and found R^2 values of 0.59, 0.81, and 0.83, respectively.[5] Our results suggests that C0 is not as good a surrogate of the AUC, contrary to what was repeatedly claimed (e.g., Shipkova et al. TDM 2016[7]), and, in this respect, no better than for tacrolimus (Brunet et al. TDM 2019[22]). Consequently, adjusting everolimus dose on the AUC rather than on C0, at least in certain patients or situations, might improve patient outcome.

One strength of the present study is to have trained and validated ML algorithms of everolimus exposure prediction on two independent datasets, both larger than those generally used in PK studies. In 2012, Moes et al.[16] trained a PK model on 52 PK profiles and developed an MAP-BE with a 2-point LSS (C0 and C2) yielding R^2 = 0.90 in the training step, but it was not validated in an independent validation dataset. Similarly, Robertsen et al. proposed a PK model to describe everolimus PK in whole blood and in peripheral blood mononuclear cells,[8] with a slightly better performance than our model trained on patient data (RMSE = 9.9% and 10.6%, respectively), but their Bayesian estimator was trained on a very small dataset (n = 20) and validated on an even smaller one (n = 4). In the study by Zwart et al.,[23] a model was built using only one observed concentration (C0) to estimate the $AUC_{0\text{-}12}$, based on 322 PK profiles. The normalized RMSE was 17.4% and 16.3% at less than or equal to 6 and greater than or equal to 6 months post-transplant, respectively. However, because their data were collected from routine clinical care, their reference AUCs were calculated using less samples for each profile (n = 4, up to 7 in the best case) than in our study (n = 10–12).

Despite training on a rather large set of data (as compared with previous literature reports), the performance of our XGBoost algorithm was just slightly better than that of our previous MAP-BE. We even observed some absolute errors >20% in the validation dataset. These atypical cases were either overestimated because of unusual flat profiles or underestimated because of very high peak concentrations at 1 h (see Figure S2). These situations were probably better covered by the simulation approach, anticipating a large number of possibilities in an artificial way (variability in doses, clearances, ideal body weights...). Figure 2 clearly shows improved predictions when AUC greater than 200 μg*h/L. Because we receive many less requests for everolimus $AUC_{0\text{-}12h}$ estimation on our ISBA expert system than for tacrolimus or mycophenolic, it would take many more years to reach the same amount of actual patient data to improve the ML algorithms and reach the same level of accuracy and precision as we did for these two drugs.[13,14] For this reason, we chose to enhance ML by extending the training dataset through mixing patient and simulated profiles, or by relying only on a large number of simulations, as a test that could be extrapolated to other drugs. The last strategy was found to be the most powerful because it yielded very good performances across the whole range of AUC values in the external validation dataset.

The predictions obtained in the current work with ML are of similar quality to those obtained with MAP-BE. The latter only used the morning dose, three concentrations, and their respective times to provide an excellent estimation of the full trapezoidal $AUC_{0\text{-}12h}$, whereas the XGBoost algorithm trained on actual patient data used two more predictors, the time elapsed between transplantation and everolimus blood sampling, and patient age. However, these additional predictors presumably did not add much information to estimate $AUC_{0\text{-}12h}$, because algorithms built exclusively from simulated data without such covariates performed as well or better. In the context of MAP-BE, this has been initially described years ago by Sheiner et al.[24] for digoxin: the addition of covariates does not carry as much information as one additional plasma concentration. This is confirmed by the excellent prediction performance in the independent patient dataset, where age and post-transplantation time were not available. Moreover, long before this work, the MAP-BE used here as a comparator was built using ca. 30–40 profiles. Consequently, this study also illustrates a fundamental difference between XGBoost and MAP-BE: data-driven algorithms cannot be any better than the data available (e.g., here, the training set included very few patient PK profiles with AUC values <50 or >200 μg*h/L) whereas compartmental PK models, if well designed, are expected to be valid even beyond values used to develop them.

We trained XGBoost ML algorithms on a large number of everolimus PK profiles simulated using a PopPK model from the literature and obtained better results than algorithms trained on a 10-fold smaller database of patient data, or on mixed databases of patient and simulated PK data. XGboost estimation based on three concentration-time points only (no other covariate) provided accurate and precise estimation of everolimus interdose AUC in a

large independent dataset of everolimus full PK profiles from kidney graft recipients. These algorithms can be used as alternatives to our previously developed Bayesian estimator available through our ISBA expert system (https://abis.chu-limoges.fr/login) and will soon be implemented in a dedicated web interface (for research purposes only), together with the recently published ML algorithms for tacrolimus and mycophenolate mofetil.

AUTHOR CONTRIBUTIONS

M.L. and P.M. wrote the manuscript. J.-B.W., P.M., and M.L. designed the research. J.D. designed one of the modeling programs. M.L. and J.-B.W. trained the algorithms. M.L., P.M., and J.-B.W. analyzed the data.

CONFLICTS OF INTEREST

The authors declared no competing interests for this work.

ORCID

Marc Labriffe https://orcid.org/0000-0001-5840-8904
Jean-Baptiste Woillard https://orcid.org/0000-0003-1695-0695
Jean Debord https://orcid.org/0000-0002-3309-1100
Pierre Marquet https://orcid.org/0000-0001-7698-0760

REFERENCES

1. Schuler W, Sedrani R, Cottens S, et al. SDZ RAD, a new rapamycin derivative: pharmacological properties in vitro and in vivo. *Transplantation*. 1997;64(1):36-42. doi:10.1097/00007890-199707150-00008
2. Nashan B, Curtis J, Ponticelli C, et al. Everolimus and reduced-exposure cyclosporine in de novo renal-transplant recipients: a three-year phase II, randomized, multicenter, open-label study. *Transplantation*. 2004;78(9):1332-1340. doi:10.1097/01.tp.0000140486.97461.49
3. Vitko S, Tedesco H, Eris J, et al. Everolimus with optimized cyclosporine dosing in renal transplant recipients: 6-month safety and efficacy results of two randomized studies. *Am J Transplant*. 2004;4(4):626-635. doi:10.1111/j.1600-6143.2004.00389.x
4. Kirchner GI, Meier-Wiedenbach I, Manns MP. Clinical pharmacokinetics of everolimus. *Clin Pharmacokinet*. 2004;43(2):83-95. doi:10.2165/00003088-200443020-00002
5. Chan L, Hartmann E, Cibrik D, Cooper M, Shaw LM. Optimal everolimus concentration is associated with risk reduction for acute rejection in de novo renal transplant recipients. *Transplantation*. 2010;90(1):31-37. doi:10.1097/TP.0b013e3181de1d67
6. Kovarik JM, Kahan BD, Kaplan B, et al. Longitudinal assessment of everolimus in de novo renal transplant recipients over the first post-transplant year: pharmacokinetics, exposure-response relationships, and influence on cyclosporine. *Clin Pharmacol Ther*. 2001;69(1):48-56. doi:10.1067/mcp.2001.112969
7. Shipkova M, Hesselink DA, Holt DW, et al. Therapeutic drug monitoring of Everolimus: a consensus report. *Ther Drug Monit*. 2016;38(2):143-169. doi:10.1097/FTD.0000000000000260
8. Robertsen I, Debord J, Åsberg A, Marquet P, Woillard JB. A limited sampling strategy to estimate exposure of Everolimus in whole blood and peripheral blood mononuclear cells in renal transplant recipients using population pharmacokinetic modeling and Bayesian estimators. *Clin Pharmacokinet*. 2018;57(11):1459-1469. doi:10.1007/s40262-018-0646-5
9. Saint-Marcoux F, Woillard JB, Jurado C, Marquet P. Lessons from routine dose adjustment of tacrolimus in renal transplant patients based on global exposure. *Ther Drug Monit*. 2013;35(3):322-327. doi:10.1097/FTD.0b013e318285e779
10. Badillo S, Banfai B, Birzele F, et al. An introduction to machine learning. *Clin Pharmacol Ther*. 2020;107(4):871-885. doi:10.1002/cpt.1796
11. Woillard JB, Salmon Gandonnière C, Destere A, et al. A machine learning approach to estimate the glomerular filtration rate in intensive care unit patients based on plasma Iohexol concentrations and covariates. *Clin Pharmacokinet*. 2021;60(2):223-233. doi:10.1007/s40262-020-00927-6
12. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16*. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785
13. Woillard JB, Labriffe M, Debord J, Marquet P. Tacrolimus exposure prediction using machine learning. *Clin Pharmacol Ther*. 2020;30:361-369. doi:10.1002/cpt.2123
14. Woillard JB, Labriffe M, Debord J, Marquet P. Mycophenolic acid exposure prediction using machine learning. *Clin Pharmacol Ther*. 2021;24:370-379. doi:10.1002/cpt.2216
15. Woillard JB, Labriffe M, Prémaud A, Marquet P. Estimation of drug exposure by machine learning based on simulations from published pharmacokinetic models: the example of tacrolimus. *Pharmacol Res*. 2021;167:105578. doi:10.1016/j.phrs.2021.105578
16. Moes DJAR, Press RR, den Hartigh J, van der Straaten T, de Fijter JW, Guchelaar HJ. Population pharmacokinetics and pharmacogenetics of everolimus in renal transplant patients. *Clin Pharmacokinet*. 2012;51(7):467-480. doi:10.2165/11599710-000000000-00000
17. Elmokadem A, Riggs MM, Baron KT. Quantitative systems pharmacology and physiologically-based pharmacokinetic modeling with mrgsolve: a hands-on tutorial. *CPT Pharmacometrics Syst Pharmacol*. 2019;8(12):883-893. doi:10.1002/psp4.12467
18. Templ M, Kowarik A, Meindl B. Statistical disclosure control for micro-data using the R package sdcMicro. *J Stat Softw*. 2015;67(1):1-36. doi:10.18637/jss.v067.i04
19. Denney W, Duvvuri S, Buckeridge C. Simple, automatic noncompartmental analysis: the PKNCA R package. *J Pharmacokinet Pharmacodyn*. 2015;42:S65-S107. doi:10.1007/s10928-015-9432-2
20. Schloerke B, Cook D, Larmarange J, et al. *GGally: Extension to "Ggplot2."*; 2021. Accessed April 8, 2021. https://CRAN.R-project.org/package=GGally
21. Signorell A, Aho K, Alfons A, et al. *DescTools: Tools for Descriptive Statistics*; 2021. Accessed April 8, 2021. https://CRAN.R-project.org/package=DescTools
22. Brunet M, van Gelder T, Åsberg A, et al. Therapeutic drug monitoring of tacrolimus-personalized therapy: second consensus report. *Ther Drug Monit*. 2019;41(3):261-307. doi:10.1097/FTD.0000000000000640

23. Zwart TC, Moes DJAR, van der Boog PJM, et al. Model-informed precision dosing of Everolimus: external validation in adult renal transplant recipients. *Clin Pharmacokinet*. 2021;60(2):191-203. doi:10.1007/s40262-020-00925-8
24. Sheiner LB, Beal S, Rosenberg B, Marathe VV. Forecasting individual pharmacokinetics. *Clin Pharmacol Ther*. 1979;26(3):294-305. doi:10.1002/cpt1979263294

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Labriffe M, Woillard J-B, Debord J, Marquet P. Machine learning algorithms to estimate everolimus exposure trained on simulated and patient pharmacokinetic profiles. *CPT Pharmacometrics Syst Pharmacol*. 2022;00:1-11. doi:10.1002/psp4.12810

Figure S1. Correlation matrix between everolimus AUC0-12h Bayesian estimates and potential predictors in the training (patient data) and testing datasets.

AUC, area under the curve. c0h, trough level; c1h, concentration at 1 hour; c2h, concentration at 2 hours; dev_t1h_rel, deviation from the 1-hour theoretical time; dev_t2h_rel, deviation from the 2-hours theoretical time; diff_c0_c1, concentration difference between C1 and C0; diff_c0_c2, concentration difference between C2 and C0; diff_c2_c1, concentration difference between C2 and C1; post_transplant_period, time between transplantation and tacrolimus blood concentrations.

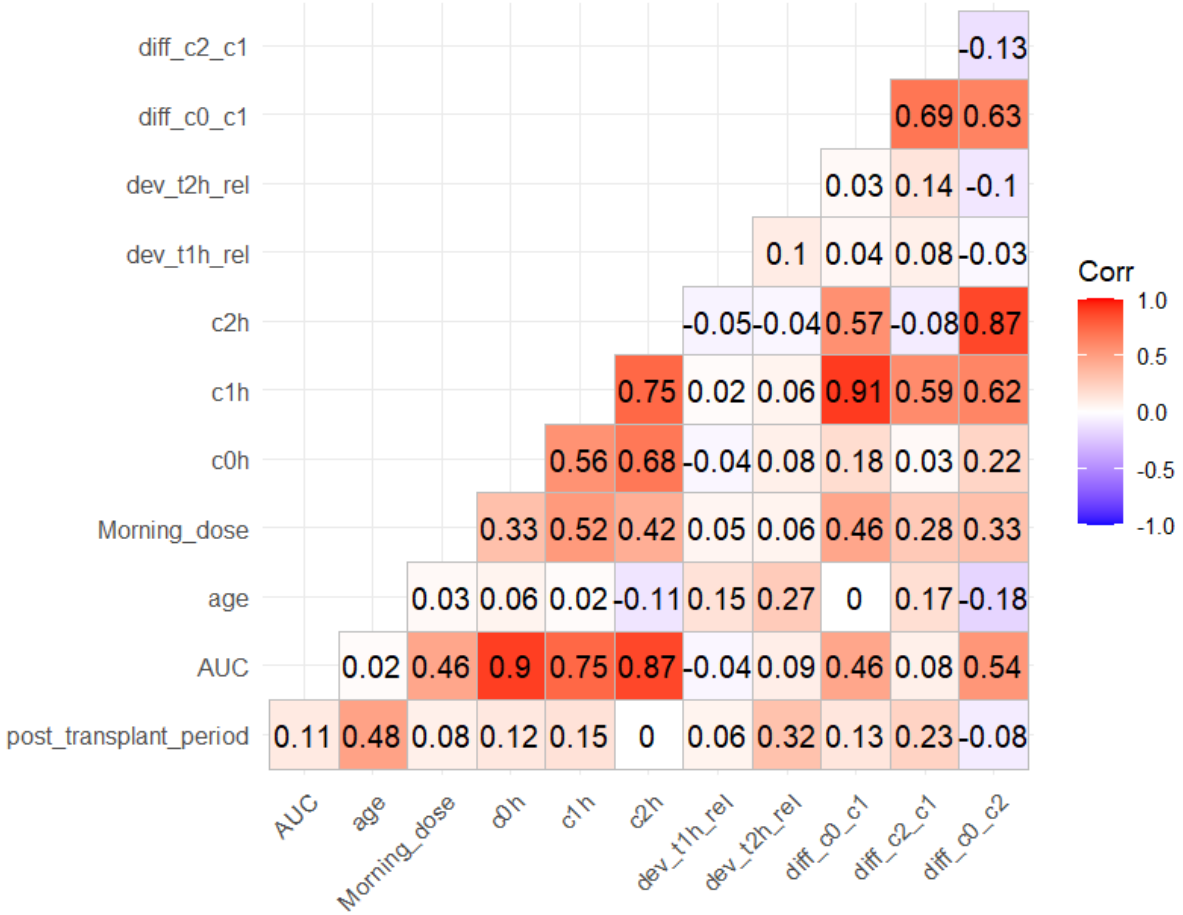


Figure S2. Pharmacokinetic profiles of the cases with absolute errors > 20 % using the best XGBoost algorithm trained on simulated PK profiles (n = 5016).

PK, pharmacokinetic; XGBoost, extreme gradient boosting, an optimized gradient boosting machine learning algorithm.

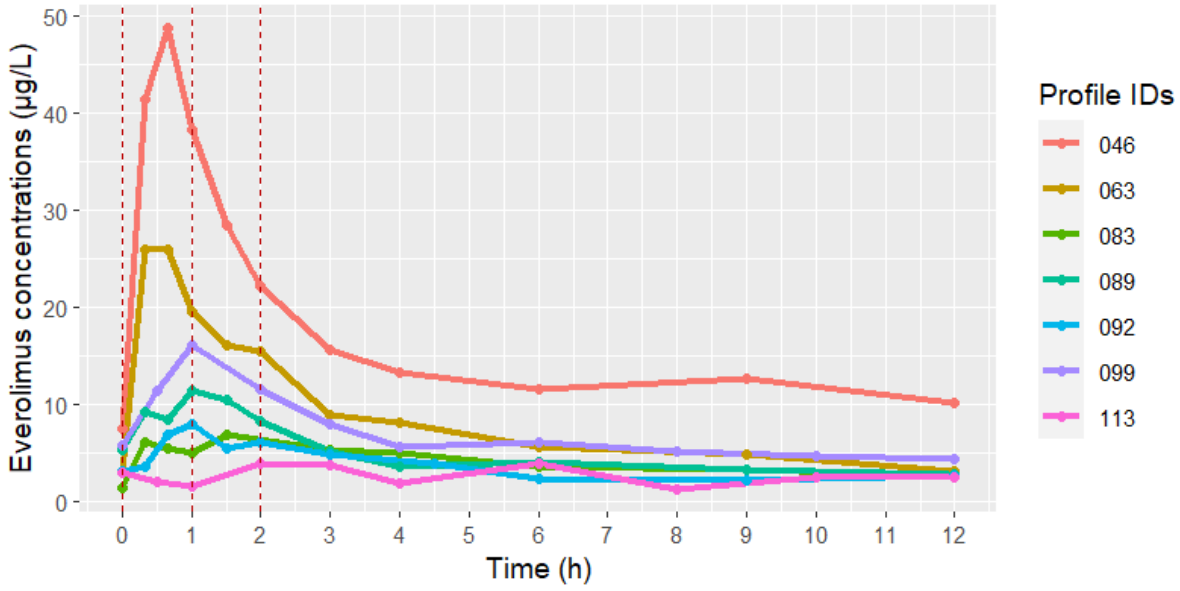


Table S1. Best-tuned XGBoost hyperparameters

	Hyperparameters	
508 patient PK profiles	mtry	2
	Trees	1000
	min_n	27
	Tree depth	2
	Learning rate	0.043
500 simulated + 508 patient PK profiles	mtry	8
	Trees	1000
	min_n	34
	Tree depth	14
	Learning rate	0.004
1003 simulated PK profiles	mtry	3
	Trees	1000
	min_n	29
	Tree depth	4
	Learning rate	0.007
1003 simulated + 508 patient PK profiles	mtry	9
	Trees	1000
	min_n	16
	Tree depth	10
	Learning rate	0.062
2508 simulated PK profiles	mtry	9
	Trees	1000
	min_n	16
	Tree depth	10
	Learning rate	0.062
2508 simulated + 508 patient PK profiles	mtry	9
	Trees	1000
	min_n	16
	Tree depth	10
	Learning rate	0.062
5016 simulated PK profiles	mtry	7
	Trees	1000
	min_n	35
	Tree depth	12
	Learning rate	0.017
5016 simulated + 508 patient PK profiles	mtry	9
	Trees	1000
	min_n	16
	Tree depth	10

	Learning rate	0.062
10,035 simulated PK profiles	mtry	9
	Trees	1000
	min_n	16
	Tree depth	10
	Learning rate	0.062
10,035 simulated + 508 patient PK profiles	mtry	9
	Trees	1000
	min_n	16
	Tree depth	10
	Learning rate	0.062

Learning rate, rate at which the boosting algorithm adapts from iteration-to-iteration; min_n, minimum number of data points in a node that are required for the node to be split further; mtry, number of predictors that will be randomly sampled at each split when creating the tree algorithms; Tree depth, maximum depth of the tree (i.e., number of splits); Trees, number of trees contained in the boosted ensemble (i.e., number of boosting iterations).

Table S2. Performance of C0 at estimating everolimus $AUC_{0\text{-}12h}$ using linear regression, trained and validated on the dataset with the full PK profiles (n = 114)

RMSE, µg*h/L	23.8
Normalized RMSE (%)	22.8
R^2	0.776
Relative MPE (%)	4.5
Number of MPE out of the ± 20 % interval: n (%)	34 (29.8 %)

AUC, area under the curve; MPE, mean prediction error; Normalized RMSE, root mean square error divided by the mean of reference AUCs; PK, pharmacokinetic; RMSE, root mean square error.

Table S3. Performance of the XGBoost algorithm with increased gaussian noise on the sampling times of C1 and C2 in the simulated training set

		gaussian noise with mean = 6.2 % and sd = 20 % added to sampling times		gaussian noise with mean = 1.2 % and sd = 3.9 % added to sampling times	
		Train set (n = 75 %)	Test set (n = 25 %)	External validation set (n = 114 full PK profiles)	
5016 simulated PK profiles	RMSE, µg*h/L	10.0	19.3	11.9	10.8
	Normalized RMSE (%)	6.6	12.3	11.4	10.3
	R^2	0.992	0.976	0.944	0.956
	Relative MPE (%)	0.1	0.0	1.1	1.6
	Number of MPE out of the ± 20% interval: n (%)	4 (0.1 %)	4 (0.3 %)	8 (7.0 %)	7 (6.1 %)

AUC, area under the curve; MPE, mean prediction error; Normalized RMSE, root mean square error divided by the mean of reference AUCs; PK, pharmacokinetic; RMSE, root mean square error.

Text S1. Code for the simulation of everolimus PK profiles and XGBoost training

```
# Packages

library(tidyverse)
library(tidymodels)
library(mrgsolve)
library(sdcMicro)


# Simulations

# Loading of the .cpp file containing the model of Moes et al. (model
attached at the end of the document)
mod <- mread("model_evero_moes", modlib())

init(mod)

data <- expand.ev(ID = 1:360, amt = c(0.5, 1, 1.25, 1.5, 1.75, 2, 2.25,
2.5, 2.75, 3, 3.25, 3.5, 4, 4.5), ii = 12, ss = 1, addl = 12)

set.seed(1234)
data2 <- data %>%
  mutate(
    IBW = rtruncnorm(length(data$ID), a = 52, b = 83, mean = 68, sd = 7.5)
# IBW: ideal body weight
  )

set.seed(1234)
sim <- mod %>%
  data_set(data) %>%
  Req(CP) %>%
  mrgsim(end = 12, delta = 0.1)

sim2 <- as_tibble(sim) %>%
  left_join(data2, by ="ID") %>%
  dplyr::rename(time = time.x)

set.seed(1234)
evero_noise <- addNoise(sim2, variables = c("time"), noise = 1)


# Training

# Data splitting

set.seed(1234)
evero_split <- initial_split(evero_ml)
evero_ml_train <- training(evero_split)
evero_ml_test <- testing(evero_split)
```

```
# Pre processing

evero_ml_rec <- recipe(AUC ~ ., data = evero_ml_train) %>%
  step_dummy(all_nominal())
evero_prep_recipe <- prep(evero_ml_rec)
evero_train_recipe <- juice(evero_prep_recipe)
evero_test_recipe <- bake(evero_prep_recipe, new_data = evero_ml_test)

# Model & workflow

# model
xgb_spec <- boost_tree(mode = "regression",
                       mtry = tune(),
                       trees = 1000,
                       min_n = tune(),
                       tree_depth = tune(),
                       learn_rate = tune()) %>%
  set_engine("xgboost")

# workflow model + recipe
xgb_wf <- workflow() %>%
  add_recipe(evero_ml_rec) %>%
  add_model(xgb_spec)

# hyperparameters
set.seed(1234)
evero_folds <- vfold_cv(evero_ml_train) # par defaut 10 fois

set.seed(1234)
tune_xgb <- tune_grid(
  xgb_wf,
  resamples = evero_folds,
  grid = 30
  )

# best model
best_rmse_xgb <- select_best(tune_xgb, "rmse",  maximize = F)

final_xgb <- finalize_model(
  xgb_spec,
  best_rmse_xgb
  )

# finalize workflow
final_wf_xgb <- workflow() %>%
  add_recipe(evero_ml_rec) %>%
  add_model(final_xgb)

# resample
set.seed(1234)
folds <- vfold_cv(evero_ml_train)
```

```
set.seed(1234)
xgb_rs <- fit_resamples(object = final_wf_xgb,
                        resamples = folds,
                        control = control_resamples(verbose = T, save_pred
= T))

# perf resample
xgb_rs %>% collect_metrics()

# fit workflow
fit_workflow <- fit(final_wf_xgb, evero_ml_train)


# test set

set.seed(1234)
final_res <- final_wf_xgb %>%
  last_fit(evero_split)

final_res %>% collect_metrics()

final_res_last <- final_res %>%
  collect_predictions()
```

Model from Moes et al. (.cpp file) :

```
[PROB]
# Population pharmacokinetics of everolimus in stable renal transplant
recipients following oral administration

https://link.springer.com/article/10.2165/11599710-000000000-00000

[SET] end=10*12, delta=0.1

[PARAM] @annotated
TVALAG : 0.709 : lagtime
TVF    :     1 : fixed bioavailability
TVCL   :  17.9 : Typical value of clearance (L/h)
TVV1   :   148 : Typical central volume of distribution (L)
TVQ    :  55.7 : Typical intercomp clearance 1 (L/h)
TVV2   :   498 : Typical peripheral volume of distribution 1 (L)
TVKA   :  7.55 : Typical absorption rate constant(1/h)
DOCL   : 0.532 : effect of dose on cl
IBV1   : -1.41 : effect of ideal body weight on v1
IBW    : 65.75 : ideal body weight

[CMT] @annotated
DEPOT : Dosing compartment (mg)
CENT  : Central comapartment (mg)
PERI  : First peripheral compartment (mg)
AUC   : AUC cp
```

```
[OMEGA] @annotated
ETACL  : 0.262 : ETA on clearance
ETAV1  : 0.277 : ETA on V1
ETAKA : 1.086 : ETA on KA


[MAIN]
double CL = TVCL*exp(ETACL);
double V1 = TVV1*exp(ETAV1)*pow(65.75/IBW,IBV1);
double KA = TVKA*exp(ETAKA);
double F = TVF;
double Q = TVQ;
double V2 = TVV2;
double ALAG = TVALAG;

[SIGMA] @annotated
PROP : 0.0001 : proportionnal Residual unexplained variability

[ODE]
dxdt_DEPOT = -KA*DEPOT;
dxdt_CENT = KA*DEPOT -(CL+Q)*CENT/V1 + Q*PERI/V2 ;
dxdt_PERI = Q*CENT/V1 - Q*PERI/V2 ;
dxdt_AUC = CENT/(V1/1000);
if(SS_ADVANCE) dxdt_AUC = 0;

[TABLE]
capture CP = (CENT/(V1/1000))*(1+PROP);
```

III. Améliorer la définition d'un critère d'efficacité important, le rejet du greffon

III.1. Objectifs

Actuellement, le diagnostic de rejet est posé à partir de l'examen histologique d'une biopsie du greffon par un spécialiste anatomopathologiste. Il repose sur l'analyse systématique des lésions au niveau de l'interstitium, des tubes, des glomérules et des structures vasculaires, permettant d'établir des scores semi-quantitatifs allant de 0 à 3. La classification de Banff est obtenue par l'application de règles complexes pour combiner ces scores et attribuer les diagnostics correspondants. L'utilisation de ces règles est malheureusement peu reproductible, certaines ne font pas consensus, et elles ne prennent pas en compte des critères cliniques potentiellement pertinents. Enfin, dans les essais cliniques qui utilisent ce critère de jugement, il peut exister une variabilité intercentre qui dessert les comparaisons entre stratégies thérapeutiques.

Dans cette étude, les objectifs étaient de : (i) construire un outil de classification des biopsies robuste, se basant sur les diagnostics consensuellement posés par un groupe d'experts internationaux (et non pas sur la classification de Banff elle-même) ; (ii) proposer une aide pour les rares cas incertains des algorithmes (zone grise) ; (iii) comparer les performances des modèles, à la stricte application de la classification de Banff ; (iv) proposer des outils améliorant l'interprétabilité des résultats fournis par l'algorithme de ML.

La base de données d'apprentissage incluait 631 biopsies provenant de l'étude Biomargin.[11] La méthode utilisée en ML était XGBoost. Les variables utilisées comme prédicteurs avaient été recueillies au moment de chaque biopsie : les scores de Banff catégorisés de 0 à 3 (g, ptc, C4d, cg, v, i, t, ti, ct, ci, cv), la présence ou non d'anticorps anti-donneur (DSA), le temps écoulé depuis la transplantation, la créatininémie, et la protéinurie. A la fin, les modèles ont été validés sur des jeux de données indépendants provenant de trois centres en Europe : Leuven, Belgique (n = 3744) ; Hanovre, Allemagne (n = 589), et Necker, Paris (n = 360).

III.2. Discussion

Les deux principaux types de rejet sont ceux médiés par les anticorps (ABMR) et ceux médiés par les cellules T (TCMR). L'expression de ces rejets peut transparaitre sur différents plans : lésions histologiques, fonction du greffon, anticorps anti-HLA du donneur (DSA), et certains biomarqueurs (C4d au niveaux de la biopsie, protéines urinaires [136]...). Ces indices sont rarement tous *détectés* à la fois, bien que certains soient liés. La classification de Banff essaye donc de prendre en compte différentes combinaisons de ces signes pour déterminer les diagnostics de rejet. Au fil du temps, les règles de Banff ont gagné en complexité, à un point tel que les humains sont devenus incapables ou peu disposés à les suivre à la lettre, ou à s'adapter à chaque nouvelle version (tous les deux ans) [137,138]. Dans cet Article 3, nous avons entraîné un algorithme à attribuer les différents diagnostics de rejet de la même manière

[11] Biomargin : BIOMArkers of Renal Graft INjuries [135].

qu'un groupe d'experts cliniciens et anatomopathologistes, en incluant quelques données cliniques.

Ce sont 631 biopsies, accompagnées de quelques informations sur leur contexte clinique et biologique, qui ont servi à entrainer les algorithmes XGBoost. Les modèles apprenaient ainsi à rendre un score entre 0 et 1 pour chaque forme de rejet. Les algorithmes étaient alors ensuite testés sur 3 cohortes indépendantes (n = 4693) avec des AUC de courbes ROC supérieures à 0,90 et des performances des courbes PR également très satisfaisantes (Article 3).

Tableau 2 : Résumé des résultats des validations externes de l'Article 3 [139]

ABMR, rejet médié par les anticorps ; PR, courbe precision (valeur prédictive positive) recall (sensibilité) ; TCMR, rejet médié par les cellules T.

		ABMR			TCMR		
Courbe	**Centre**	**AUC**	**Minimum (classification au hasard)**	**Maximum atteignable**	**AUC**	**Minimum (classification au hasard)**	**Maximum atteignable**
ROC	Leuven	0,97	0,5	1	0,94	0,5	1
	Hanovre	0,97	0,5	1	0,94	0,5	1
	Paris Necker	0,95	0,5	1	0,91	0,5	1
PR	Leuven	0,72	0,07	1	0,83	0,18	1
	Hanovre	0,92	0,06	1	0,91	0,33	1
	Paris Necker	0,84	0,24	1	0,55	0,13	1

Quand les résultats de l'algorithme étaient comparés aux diagnostics faits dans les centres de référence, les modèles de ML étaient plus souvent en accord avec leur conclusion que la classification de Banff elle-même (Tableau 3 de l'article). Dans l'éditorial du numéro de l'*American Journal of Transplantation* dans lequel a été publié l'Article 3, l'actuel président de la fondation Banff admettait que cet outil de ML permettait de mieux capturer le raisonnement intuitif d'un médecin que l'application exacte des règles de Banff [137].

Au final, cette étude a permis de montrer que l'IA peut apprendre de l'expertise humaine, pour aider en retour des centres de transplantation plus petits, avec des anatomopathologistes moins expérimentés.

En ce qui concerne la fibrose interstitielle et l'atrophie tubulaire (IFTA), nous voulions également proposer un outil robuste d'interprétation. Les AUC ROC et PR très proches de 1 dans les trois jeux de validation (Figure 3 de l'article) suggèrent que ce diagnostic est déjà très reproductible d'un centre à l'autre. Nous avons également constaté que, assez logiquement,

les lésions ct et ci étaient essentielles : ci correspond au pourcentage de la surface du cortex qui est fibreuse, et ct correspond au pourcentage de surface de cortex contenant des tubes atrophiques. Les deux scores sont ceux utilisés pour le diagnostic IFTA dans la classification de Banff 2015. En général, c'était le score ct qui était décisif (Figure 2 de l'article). Hormis peut-être la protéinurie, les autres variables (lésions élémentaires de rejet, créatininémie...) n'étaient pas aussi fortement liées à ce diagnostic. Dans ce contexte, il ne nous est pas apparu utile de le comparer à la classification de Banff, car cette catégorie fait probablement déjà consensus.

Un modèle de gradient boosting comme ceux utilisés dans cet article est constitué de plusieurs centaines d'arbres successifs (voir Les méthodes par arbres). Bien qu'il puisse être facile d'utilisation sur le logiciel R, ce type de modèle ne peut être représenté aisément comme pourrait l'être un arbre de décision seul, ou une équation de régression logistique. Il faut donc pouvoir être capable de comparer l'importance des variables et d'expliquer le résultat de chaque prédiction. C'est ce que nous avons essayé de faire en présentant l'importance des variables dans la Figure 2 de l'article, c'est-à-dire la capacité globale d'une variable à séparer les *rejets* des *non-rejets* à chaque fois qu'elle est utilisée dans un nœud. On parle d'*intelligibilité globale*. De plus, nous savons qu'à l'avenir les valeurs SHAP ou des équivalents seront indispensables à fournir avec *chaque prédiction* (Figure 3 de l'article) : c'est l'*intelligibilité locale*. Ici, les valeurs SHAP permettaient pour un greffon donné de quantifier l'impact de chaque information disponible (et son sens), sur la prédiction qui lui était associée par le modèle.

Bien sûr, dans cette approche, le résultat du modèle dépend de la lecture et de la quantification des lésions élémentaires en amont, par un humain. Ces scores sont en effet gradés de 0 à 3 par l'anatomopathologiste, en fonction de l'étendue des lésions, exprimé en pourcentage d'une aire (cas le plus fréquent). Cette évaluation humaine peut être sujette à variabilité, malgré des aides existantes en ligne [140]. Ici, nos classificateurs améliorent l'*interprétation* des lésions élémentaires et des tests de laboratoire associés, en supposant qu'ils sont exacts. Lorsque ce n'est pas le cas, l'interprétation peut ne pas être exacte non plus.

Dans ce travail, nous avons mis en évidence des différences entre les diagnostics de la classification de Banff, et ceux des modèles (qui copient les experts). Ces différences observées sont-elles cliniquement pertinentes ? [137] Pour pouvoir le vérifier, nous pourrions mettre en parallèle, pour les cas divergents, les diagnostics de chacun des deux systèmes et le devenir à long terme des patients (date de la perte de fonction du greffon). De cette manière nous serions capables de visualiser les différences de survie du greffon grâce aux dates de retour en dialyse, à l'aide de courbes de Kaplan-Meier.

III.3. Article 3

Machine learning-supported interpretation of kidney graft elementary lesions in combination with clinical data

Am J Transplant. 2022

Received: 8 February 2022 | Revised: 22 August 2022 | Accepted: 25 August 2022

DOI: 10.1111/ajt.17192

ORIGINAL ARTICLE

Machine learning-supported interpretation of kidney graft elementary lesions in combination with clinical data

Marc Labriffe[1,2] | Jean-Baptiste Woillard[1,2] | Wilfried Gwinner[3] | Jan-Hinrich Braesen[4] | Dany Anglicheau[5,6,7] | Marion Rabant[8] | Priyanka Koshy[9] | Maarten Naesens[10,11] | Pierre Marquet[1,2]

[1]Pharmacology & Transplantation, INSERM U1248, Université de Limoges, Limoges, France

[2]Department of Pharmacology, Toxicology and Pharmacovigilance, CHU de Limoges, Limoges, France

[3]Nephrology, Internal Medicine, Hannover Medical School, Hannover, Germany

[4]Institute for Pathology, Nephropathology Unit, Hannover Medical School, Germany

[5]Université de Paris, Paris, France

[6]INSERM U1151, Paris, France

[7]Department of Nephrology and Kidney Transplantation, Necker Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France

[8]Department of Pathology, Necker Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France

[9]Department of Pathology, University Hospitals Leuven, Leuven, Belgium

[10]Nephrology and Renal Transplantation Research Group, Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium

[11]Department of Nephrology and Renal Transplantation, University Hospitals Leuven, Leuven, Belgium

Correspondence
Pierre Marquet, Department of Pharmacology & Transplantation, INSERM U1248, Université de Limoges, Limoges, France.
Email: pierre.marquet@unilim.fr

Interpretation of kidney graft biopsies using the Banff classification is still heterogeneous. In this study, extreme gradient boosting classifiers learned from two large training datasets (n = 631 and 304 cases) where the "reference diagnoses" were not strictly defined following the Banff rules but from central reading by expert pathologists and further interpreted consensually by experienced transplant nephrologists, in light of the clinical context. In three external validation datasets (n = 3744, 589, and 360), the classifiers yielded a mean ROC curve AUC (95%CI) of: 0.97 (0.92–1.00), 0.97 (0.96–0.97), and 0.95 (0.93–0.97) for antibody-mediated rejection (ABMR); 0.94 (0.91–0.96), 0.94 (0.92–0.95), and 0.91 (0.88–0.95) for T cell–mediated rejection; >0.96 (0.90–1.00) with all three for interstitial fibrosis–tubular atrophy. We also developed a classifier to discriminate active and chronic active ABMR with 95% accuracy. In conclusion, we built highly sensitive and specific artificial intelligence classifiers able to interpret kidney graft scoring together with a few clinical data and automatically

Abbreviations: ABMR, antibody-mediated rejection; ah, arteriolar hyalinosis; AI, artificial intelligence; AUC, area under the curve; BIOMARGIN, BIOMArkers of Renal Graft INjuries; BKVN, BK virus nephropathy; C4d, linear C4d staining in ptc or medullary vasa recta; cg, chronic transplant glomerulopathy; ci, interstitial fibrosis in cortex; ct, tubular atrophy in cortex; cv, arterial intimal fibrosis (fibrointimal thickening); DSA, donor-specific antibodies; g, glomerulitis; i, inflammation in non-scarred cortex; IFTA, interstitial fibrosis/tubular atrophy grade II; IQR, interquartile range; MD, missing data; ML, machine learning; NA, not applicable; Normal, refers to cases with no graft alterations; NPV, negative predictive value; PPV, positive predictive value; Precision, positive predictive value; ptc, peritubular capillaritis; PVN, polyomavirus nephropathy; Recall, sensitivity; ROC curve, receiver operating characteristic curve; ROCKET, Reclassification using OmiCs integration in KidnEy Transplantation; SHAP, SHapley Additive exPlanations; t, tubulitis in cortical tubules within non-scarred cortex; TCMR, T cell–mediated rejection; ti, total cortical inflammation; v, endarteritis (intimal arteritis); Youden Index, sensitivity + specificity - 1.

diagnose rejection, with excellent concordance with the Banff rules and reference diagnoses made by a group of experts. Some discrepancies may point toward possible improvements that could be made to the Banff classification.

KEYWORDS
biopsy, classification systems: Banff classification, clinical research / practice, kidney transplantation / nephrology, rejection: antibody-mediated (ABMR), rejection: T cell mediated (TCMR)

1 | INTRODUCTION

The international Banff classification standardizes the diagnosis of kidney allograft rejection.[1] Antibody-mediated rejection (ABMR), T cell–mediated rejection (TCMR), and "others" are inferred from combinations of histological lesions.[1] However, the interobserver reproducibility of reporting and ranking histological lesions is suboptimal.[2–5] Plasma DSA, C4d staining, serum creatinine, proteinuria, and other routine laboratory test results or clinical data may influence pathologists' or clinicians' judgment of a case. Moreover, the definition of the phenotypes has dynamically evolved with each revision of the Banff classification since 2005.[1] In small transplant centers where pathologists have a general practice, it is challenging to integrate these frequent updates with the same level of expertise as pathologists specialized in the analysis of kidney allograft biopsies. Also, in clinical studies, centralized or consensual biopsy assessment by specialized pathologists is not always possible due to logistic or financial constraints.

Three strategies may be used for assessing ABMR, TCMR, or other biopsy lesions. The first is to apply the Banff rules strictly and automatically using "if then else" operations. The second is to identify clusters of elementary lesions in an unsupervised manner, which cannot be directly compared with the Banff reference classes and whose pertinence is generally evaluated based on further patient outcome.[6] The third strategy, as yet unexplored, is to identify case clusters in a supervised manner based on "reference diagnoses," so as to automate and homogenize biopsy interpretation. Machine learning (ML) is a subset of artificial intelligence (AI) that groups interpretation or prediction algorithms capable of automatically learning and continuously adapting. With sufficient data, it can handle noisy and correlated variables, sometimes without the need for parametric assumptions, contrary to most traditional statistics. As recognized at the last Banff Meeting,[1] the combination of quality and quantity of input data is key for achieving accurate results when using ML.

The aim of the present study was to build robust and accurate ML classifiers able to hierarchize and finely combine the Banff criteria and clinical data used by pathologists and clinicians to infer TCMR, ABMR (overall, active, and chronic active), and IFTA (although it is no more a Banff category in itself). The classifiers were trained on two large databases of biopsy cases interpreted in light of the clinical context by a centralized group of expert pathologists and transplant physicians, as part of two European research programs. Their accuracy in assigning the right diagnoses as compared with the "reference diagnoses" was assessed in three other large datasets from different European countries, including one obtained through strict application of the Banff rules.

2 | MATERIALS AND METHODS

2.1 | Patients and biopsies

Histological and clinical data came from several independent datasets of elementary Banff scores and reference diagnoses. For the training set, we used data from two European programs, BIOMArkers of Renal Graft INjuries (BIOMARGIN, NCT02832661) and Reclassification using OmiCs integration in KidnEy Transplantation (ROCKET, funded by ERACoSysMed).[7] All the biopsies were read and interpreted locally and then sent to an independent expert pathologist for central reading. The discrepancies were adjudicated by a group of three independent expert pathologists. The consensual histological interpretations were reviewed and the final, "reference diagnoses" made considering the clinical context by a group of four transplant physicians (Figure 1).

For external validation of the ML classifiers, we used data from three transplant centers: KU Leuven (Belgium), MH Hannover (Germany), and Necker Paris (France).[8] All the patients of these cohorts were different from those included in the BIOMARGIN and ROCKET studies. Our study was approved by the respective local ethics committees.

For each biopsy, expert renal pathologists evaluated the elementary Banff criteria as recommended in the 2013 Banff Classification[9,10]: glomerulitis (g), peritubular capillaritis (ptc), linear C4d staining in ptc or medullary vasa recta (C4d), chronic transplant glomerulopathy (cg), endarteritis (intimal arteritis, v), inflammation in non-scarred cortex (i), tubulitis in cortical tubules within non-scarred cortex (t), total cortical inflammation (ti), tubular atrophy in cortex (ct), interstitial fibrosis in cortex (ci), arteriolar hyalinosis (ah), and arterial intimal fibrosis (fibrointimal thickening, cv). The databases also included donor-specific antibodies (DSA), serum creatinine (μmol/L), and proteinuria (g/L) at the time of biopsy. Active ABMR (yes/no), TCMR (yes/no, borderline cases included as yes), IFTA lesions (grade ≥ II), were diagnosed using all the information available and sometimes based on patient history (i.e., not always reflecting strict application of the Banff rules). They were considered as the "reference diagnoses" (*gold standard*) for training our

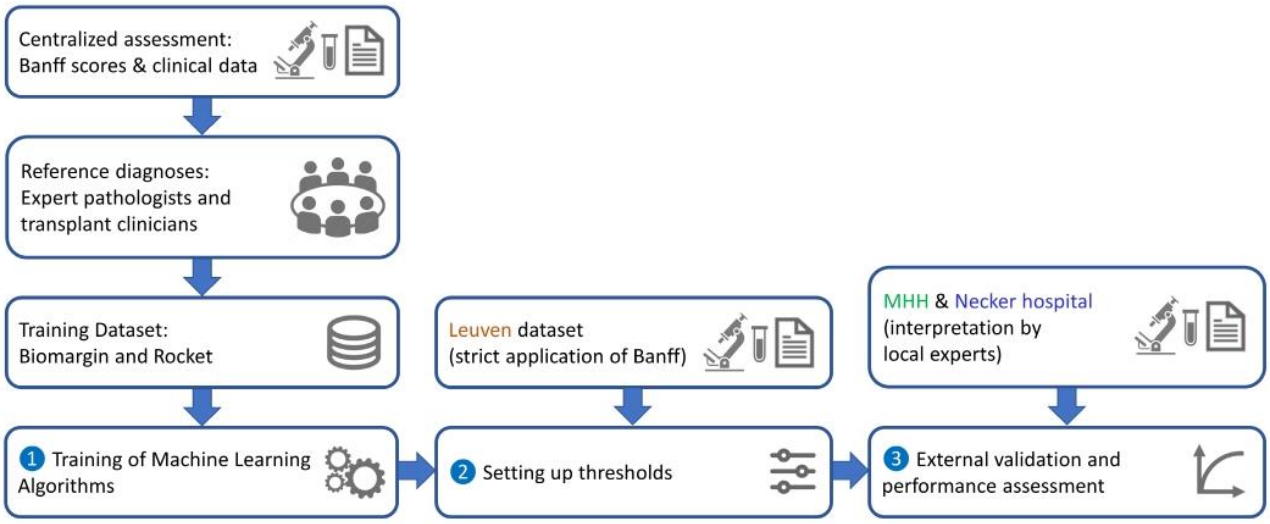


FIGURE 1 Workflow diagram of the study.

ML classifiers. No classifier was built for glomerulonephritis or polyomavirus nephropathy (PVN) because there were too few cases in the ROCKET training dataset and in the external validation datasets, but a simple "if then else" rule was applied to biopsies with positive BK viremia and positive t and i, so as to suggest using SV40 staining to avoid false positive TCMR diagnoses.

2.2 | Statistical analyses

The features considered were the Banff criteria semi-quantitatively scored from 0 to 3, DSA positivity, serum creatinine, proteinuria, and time elapsed between transplantation and biopsy. A different ML classifier was built for each outcome: active ABMR (yes/no), TCMR (yes/no), and IFTA (yes/no) using the BIOMARGIN dataset; ABMR (active/chronic active) using the ROCKET dataset. For a given case, all the results could be positive for the first three clinical phenotypes, and it is only when they were all negative that the case was presumed to be normal. The fourth classifier was only meant for ABMR-positive cases, to distinguish active from chronic active ABMR. We used Extreme Gradient Boosting (XGBoost, an ensemble method based on decision trees) which is known to perform best on structured tabular data and is able to handle missing data.[11,12] Prior to training the classifiers, we optimized the hyperparameters using 10-fold cross validation, for best accuracy. With this optimal set of hyperparameters (Table S1), we also assessed performance in the training phase as part of 10-fold cross validation. Receiver operating characteristic (ROC) curves were used to assess the threshold-independent classification performance of each model. As the training dataset was imbalanced (skewed toward normal biopsies), we also used precision–recall (PR) curves[13,14] to represent precision (positive predictive value [PPV]) versus recall (sensitivity), excluding true negatives. The minimum PR area under the curve (AUC) is equal to the prevalence of the disease. When thresholds were set at a certain value, accuracy was assessed by measuring the level of agreement between classifiers and experts. Our primary endpoints were the diagnostic accuracy and the ROC AUC of the different XGBoost classifiers. In order to improve the transparency of the classifiers, we used the SHAP (SHapley Additive exPlanations) method[15] that explains each output (here, each diagnosis made for each individual) by showing graphically the contribution of each feature to this output.

The KU Leuven cohort where all the diagnoses were made by strictly applying the Banff rules was used to set thresholds based on accuracy, PPV, and negative predictive value (NPV), (Figure 1).[16] We then explored the capacity of the algorithms as compared to a strict application of the Banff classification to reproduce the interpretation of the MHH and Necker expert pathologists. Sensitivity, specificity, PPV, NPV, and balanced accuracy ([sensitivity + specificity]/2) were calculated using local diagnoses as the reference.

For statistical computing and graphics, we used R (version 4.0.3), in particular the XGBoost package for classification (version 1.2.0.1).

3 | RESULTS

In the BIOMARGIN training dataset (n = 643), 12 biopsies were excluded because they missed ≥3 Banff elementary lesion scores (without any particular pattern, suggesting that many were not informed because they were nil). Four of these biopsies were initially qualified as adequate (≥10 glomeruli and ≥2 arteries), three were of poorer quality (≥8 glomeruli and ≥1 artery), four were inadequate, and we had no information about the quality of the last one. Among the remaining 631 cases, 73 biopsies missed one Banff elementary lesion score and 29 missed two. Patient characteristics at the time of biopsy are presented in Table 1 and other characteristics of the training dataset in Table S2. Among the 304 biopsies of the ROCKET dataset, none had missing data (this was a study exclusion criterion).

TABLE 1 Patient characteristics, laboratory test results at the time of allograft biopsy, and histological diagnoses

Variables	BIOMARGIN (training) ($n = 631$)	ROCKET (training) ($n = 304$)	KU Leuven (validation) ($n = 3744$)	MH Hannover (validation) ($n = 589$)	Necker Paris (validation) ($n = 360$)
Time after transplant (months), median (IQR)	12 (3–25)	12 (3–44)	12 (3–25)	4 (2–12)	12 (2–47)
Indicated biopsy, *n* (%)	222 (35.2)	134 (44.1)	979 (26.1)	MD	MD
Pathologic diagnosis					
ABMR, *n* (%)	104 (16.5)	107 (35.2)	242 (6.7)	36 (6.1)	86 (23.9)
TCMR, *n* (%)	82 (13.0)	84 (27.6)	665 (17.8)	193 (33.3)	47 (13.1)
Mixed ABMR/TCMR, *n* (%)	28 (4.4)	19 (6.2)	79 (2.1)	15 (2.5)	13 (3.6)
BKVN, *n* (%)	0 (0.0)	13 (4.3)	124 (3.3)	23 (4.1)	11 (3.1)
IFTA, *n* (%)	210 (33.3)	98 (32.9)	780 (20.8)	44 (8.2)	188 (52.2)
Normal, *n* (%)	312 (49.4)	93 (30.6)	2420 (65.9)	317 (57.3)	133 (36.9)
Laboratory test results at the time of the biopsy					
Serum creatinine (μmol/L), median (IQR)	150 (118–198)	154 (119–208)	141 (111–199)	172 (131–234)	176 (142–234)
DSA positivity, *n* (%)	124 (19.7)	87 (28.6)	299 (8.3)	11 (4.8)	142 (41.0)
Proteinuria (g/L), median (IQR)	0.10 (0.07–0.24)	0.10 (0.07–0.34)	MD	0.05 (0.04–0.10)	0.20 (0.08–0.47)

Abbreviations: ABMR, active antibody-mediated rejection; BKVN, BK virus nephropathy; DSA, donor-specific antibodies; IFTA, interstitial fibrosis/tubular atrophy grade II; IQR, interquartile range; MD, missing data; Normal, refers to cases with no graft alterations; TCMR, T cell–mediated rejection.

Detailed results of cross-validation in the training set are presented as Supplemental Information. For the four models, the contribution (so-called "importance") of the histological and clinical features is shown in Figure 2. "Importance" indicates how useful or valuable each feature is in the construction of the boosted decision trees within the model.[17]

Figure 3 shows the ROC and PR curves obtained in the three validation datasets. The ABMR classifier yielded ROC curve AUC of 0.97 (95%CI: 0.92–1.00), 0.97 (95%CI: 0.96–0.97), and 0.95 (95%CI: 0.93–0.97), and PR curve AUC of 0.92, 0.72, and 0.84 for the MH Hannover, KU Leuven, and Necker Paris datasets, respectively. In comparison, the minimum PR curve AUC for a No-Skill Classifier was, respectively, 0.06, 0.07, and 0.24. For the TCMR model, the ROC AUCs were 0.94 (95%CI: 0.91–0.96), 0.94 (95%CI: 0.92–0.95), and 0.91 (95%CI: 0.86–0.95); the PR curve AUCs (minimum AUC for a No-Skill Classifier) were 0.91 (0.33), 0.83 (0.18), and 0.55 (0.13). The IFTA model performance was even better with AUC ≥0.95 (95%CI: 0.90–1.00) for the ROC and PR curves, in all local datasets.

Figure 4 shows SHAP explanation force plots for two typical cases. Each feature is an arrow that pushes to increase or decrease the diagnosis score.

Thresholds were chosen to maximize accuracy in the KU Leuven cohort (Figure 1 and Figure 5). We opted for a "gray zone" with two numerical cutoffs constituting its borders. The first cutoff was used to exclude each type of diagnosis with near certainty (favoring sensitivity and NPV), and the second to assert them with near certainty (favoring specificity and PPV). The lower and upper thresholds were chosen at 0.10 and 0.75, respectively, for the binary models of ABMR and TCMR (Table 2). Between these two thresholds, the ABMR gray zone included 11.8%, 0.6%, and 2.1% of biopsies in the KU Leuven, MH Hannover, and Necker Paris datasets, respectively. The TCMR gray zone included 18.5%, 1.1%, and 0.9% of biopsies, respectively. For IFTA, the scores were already very well differentiated so we chose a single threshold of 0.10 (Table S3). Using an arbitrary threshold of 0.5, our final ABMR and TCMR classifiers did better than the strict application of the Banff rules to predict the pathologists' calls in the two independent datasets, except for TCMR cases in the MHH database where the performance was similar (Table 3). They did even better using the thresholds at 0.10 and 0.75 combined.

We explored the different combinations of features statistically associated with the "gray" score values, by building a regression tree (Table 4). Negative DSA, cg0, g<2, and g0 or ptc0 (in increasing order of importance) played against the diagnosis of ABMR. For TCMR, the situation is more complex and noticeably depends on the posttransplant period, serum creatinine and DSA.

We also compared the distribution of the features of the positive ABMR or TCMR cases obtained by strict application of the Banff rules, to those predicted by our algorithms (Table 5). ABMR positives had slightly higher g, ptc, cg, and serum creatinine mean values with our classifier than with Banff, and vice versa for DSA and C4d positivity. TCMR positives had slightly higher ptc, t, and i with our classifier, as opposed to slightly higher C4d and serum creatinine and above all higher v with Banff.

A detailed evaluation of the active/chronic active ABMR estimator in the KU Leuven cohort is presented in Table 6 (and in Supporting Information): 63 cases had active ABMR and 44 chronic active ABMR, accuracy was 0.98 for scores above 0.75 or 0.10 (i.e., including the gray zone). The final accuracy of the combination of the two ABMR estimators (yes/no and active/chronic active) successively applied to the KU Leuven dataset was 0.95.

We applied our classifiers to the six case-based scenarios used by Schinstock et al.[5] for their international survey among clinicians

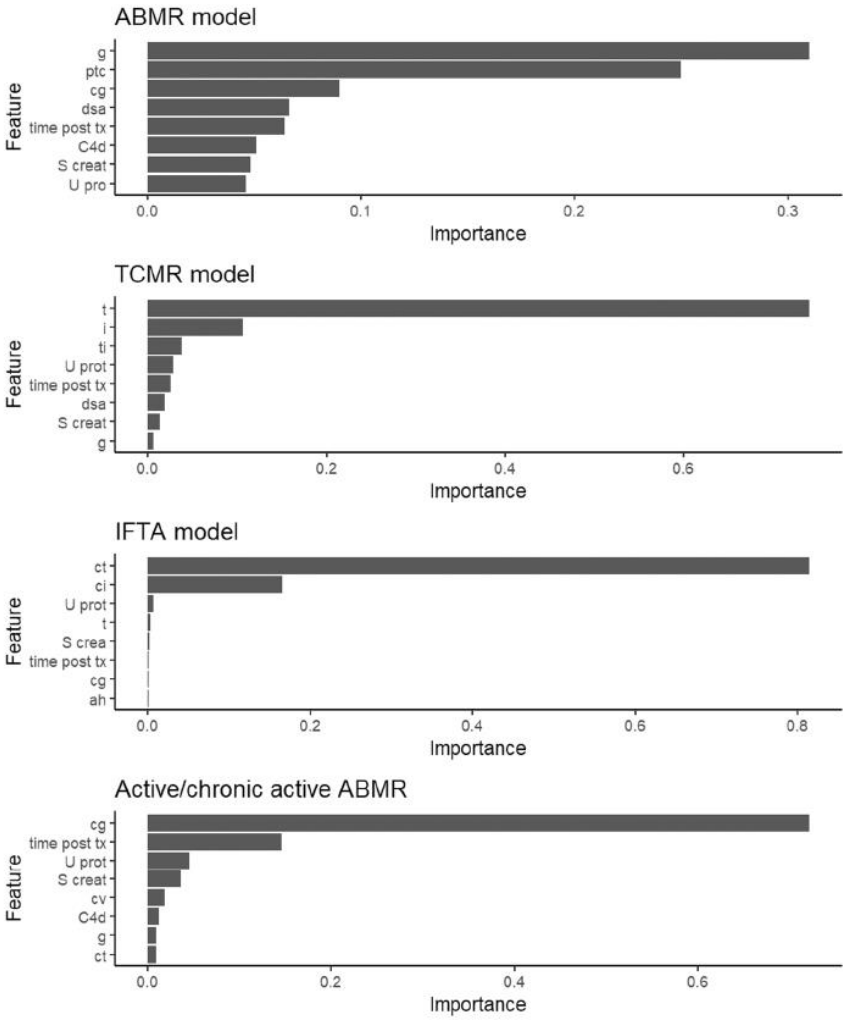


FIGURE 2 Importance of the histological and clinical features for XGBoost predictions of ABMR, TCMR, IFTA, and active/chronic active ABMR. "Importance" provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. ah, arteriolar hyalinosis; C4d, linear C4d staining in ptc or medullary vasa recta; cg, chronic transplant glomerulopathy; ci, interstitial fibrosis in the cortex; ct, tubular atrophy in the cortex; cv, arterial intimal fibrosis (fibrointimal thickening); g, glomerulitis; i, inflammation in non-scarred cortex; ptc, peritubular capillaritis; S creat, serum creatinine; t, tubulitis in cortical tubules within non-scarred cortex; ti, total cortical inflammation; dsa, donor-specific antibodies; time post tx, time after transplant; U prot, proteinuria.

and renal pathologists to understand how the Banff ABMR classification is interpreted in practice. Model predictions were perfectly consistent with the reference diagnoses (100% agreement) and without any doubt regarding the score values (Table S4).

Specific cases were also studied in detail. Mixed ABMR/TCMR cases were predicted: for 68%, 22%, and 10% of them as ABMR, gray zone, and not ABMR, respectively; for 85%, 7%, and 8% as TCMR, gray zone, and not TCMR, respectively. Among the ABMR cases, 34%, 54%, and 12% of those with negative DSA were classified as ABMR, gray zone, and not ABMR, respectively. Borderline TCMR[1] cases were all predicted as TCMR. Table S5 provides help to distinguish borderline from TCMR cases, when i and t are not available (Banff classification not applicable) but additional clinical information is.

4 | DISCUSSION

We developed AI classifiers that automatically derive the main kidney graft rejection phenotypes using the elementary histological Banff scores and a few additional laboratory test results or clinical data named "features" here. The training set was made up of two large databases of such features, together with their "reference diagnoses," that is, interpretations obtained in state-of-the art conditions. These classifiers showed excellent concordance with the diagnoses made locally by specialized pathologists and transplant physicians in independent patient cohorts from three European transplant centers.

Since ABMR, TCMR, and IFTA are not exclusive, there are eight possible combinations (including normal). Some were too rare in our training set to be able to satisfactorily train a unique classifier to sort out each combination independently. Also, the impossibility of drawing importance plots for each type of rejection would have made such a classifier look like a "Black box," less acceptable clinically. Consequently, we chose to build a separate classifier for each diagnosis, with much more data to train the algorithm and the ability to evaluate the performance of each classifier, drawing importance plots and combining the classifier results to describe mixed cases.

The IFTA grade is easily assessed using only two criteria of the Banff classification, so it is not surprising that our model almost never failed. At least, this classifier shows that no other criterion or clinical data influenced experts' decisions for this phenotype

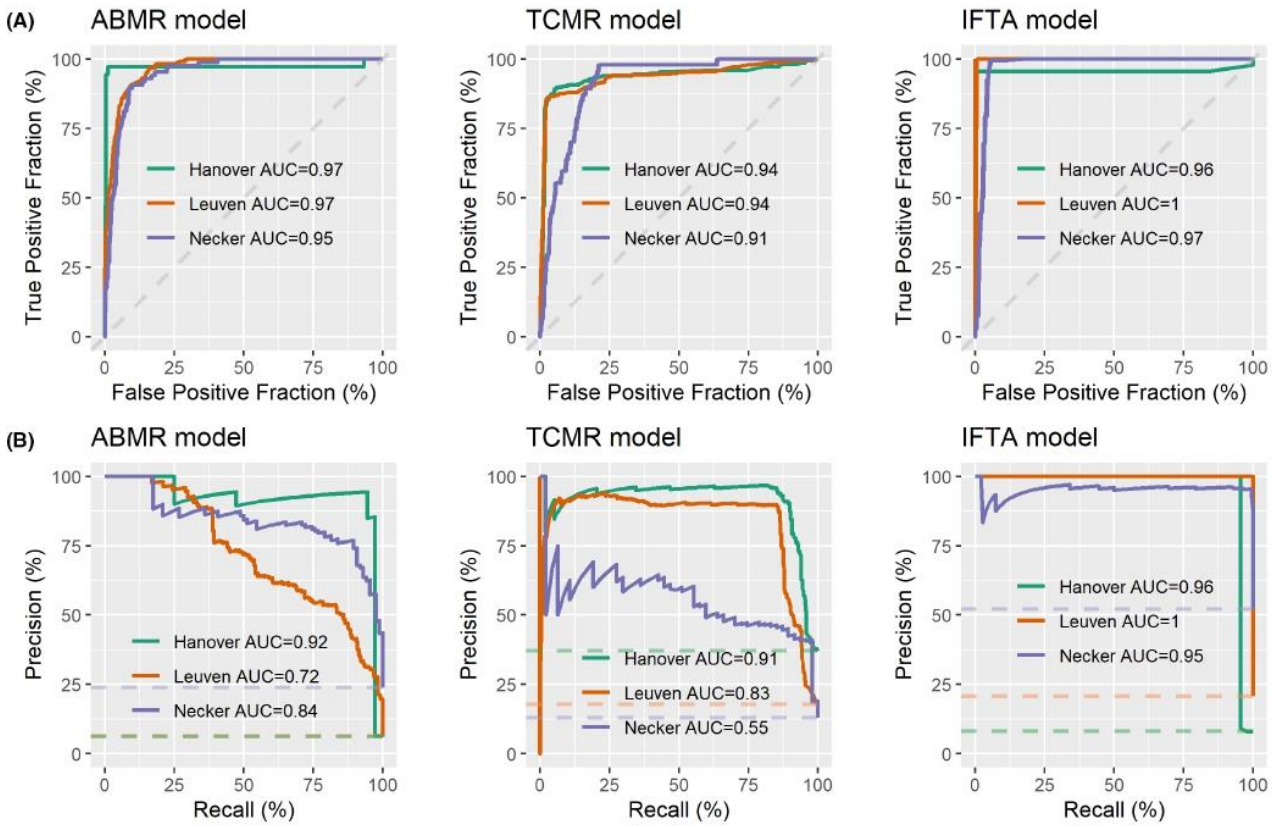


FIGURE 3 External validation of the machine learning estimators in three independent cohorts. (A) ROC curves of the ABMR model, the TCMR model, and the IFTA model. (B) PR curves of the ABMR model, the TCMR model, and the IFTA model. ABMR, active antibody-mediated rejection; IFTA, interstitial fibrosis/tubular atrophy grade II; Precision, positive predictive value; Recall, sensitivity; TCMR, T cell–mediated rejection.

FIGURE 4 SHAP explanation force plot for two exemplary cases. (A) The first case has a high score for active ABMR (0.90, which qualifies for "positive"). High serum creatinine, positive DSA, and some specific elementary lesions (C4d = 2, g = 3) increase the score. Negative proteinuria (<0.3 g/L) and some negative criteria (ptc = 0, cg = 0) decrease the score. (B) The second case has a high score for TCMR (0.93, which qualifies for "positive"). High serum creatinine, positive DSA, and some specific elementary lesions (i = 2, t = 2, ti = 3) increase the score. Negative proteinuria (<0.3 g/L) and nonspecific lesions (ct = 2) decrease the score. ABMR, active antibody-mediated rejection; C4d, linear C4d staining in ptc or medullary vasa recta; cg, chronic transplant glomerulopathy; ct, tubular atrophy in cortex; DSA, donor-specific antibodies; i, inflammation in non-scarred cortex; SHAP, SHapley Additive exPlanations; t, tubulitis in cortical tubules within non-scarred cortex; TCMR, T cell–mediated rejection; ti, total cortical inflammation.

(even if the importance of ct was higher than that of ci) and this was the only perfectly reproducible phenotype across hospitals and pathologists.

The fact that we could not predict the other phenotypes with ROC AUC = 1, despite the use of gradient boosting (an ensemble method literally based on decision trees), is probably due to poorer

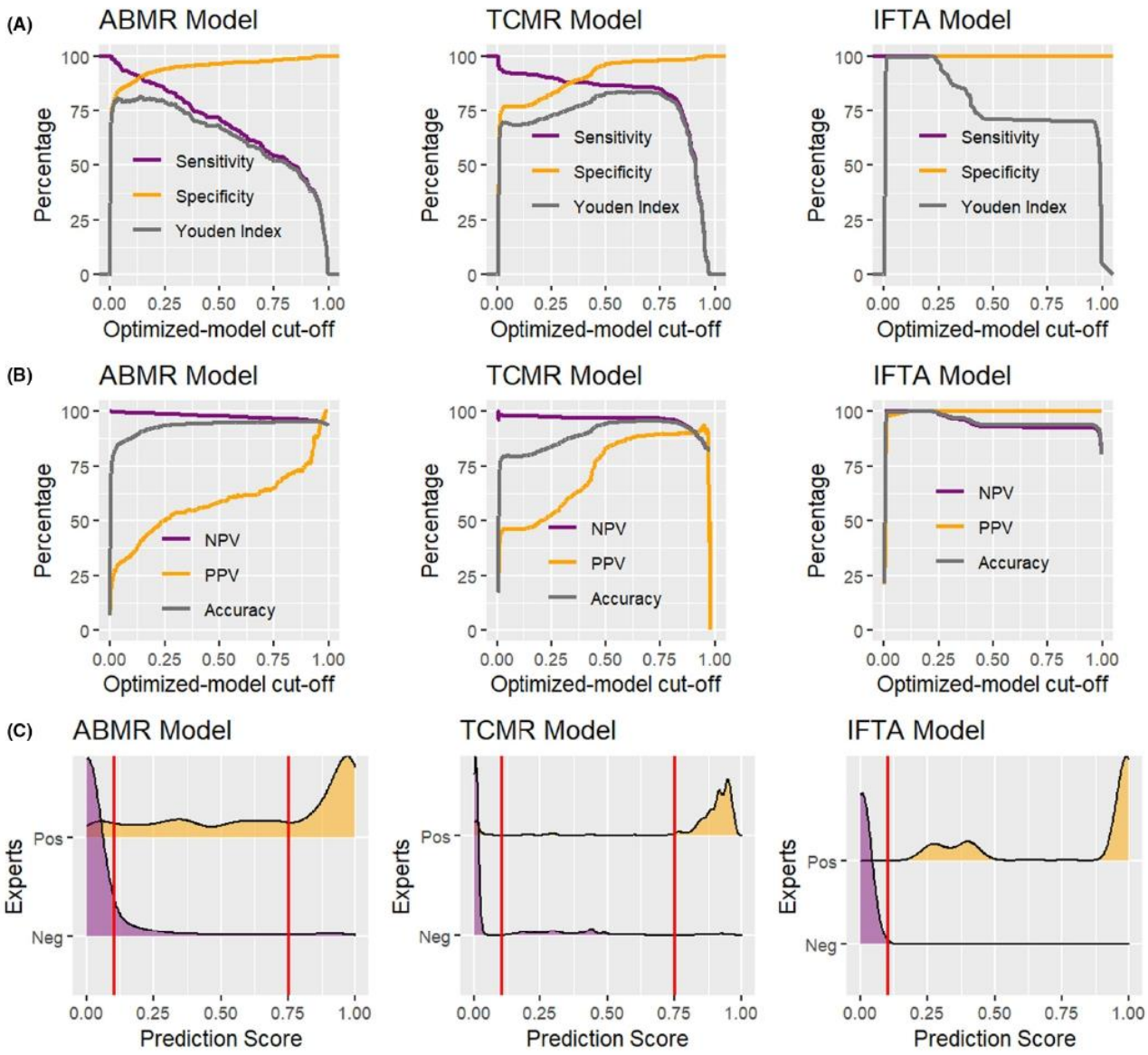


FIGURE 5 Choice of thresholds in the KU Leuven dataset. The plots at the bottom show the density of the scores. (A) Sensitivity, specificity, and Youden Index, depending on the cut-off for each model. (B) Negative predictive value, positive predictive value, and accuracy, depending on the cut-off for each model. (C) Density of the scores for each model. ABMR, active antibody-mediated rejection; IFTA, interstitial fibrosis/tubular atrophy grade II; NPV, negative predictive value; PPV, positive predictive value; TCMR, T cell–mediated rejection; Youden Index, sensitivity + specificity - 1.

reproducibility among pathologists for the interpretation of the Banff classification for these phenotypes. Surprisingly, not all predictors retained by the other classifiers after unsupervised selection were consistent with those proposed by the Banff classification. In the TCMR model, i was used much less than t, whereas in the Banff classification they are of equal importance. The most probable explanation for this is the lower frequency of t0i1+ than i0t1+ cases in the training set, and the higher percentage of TCMR diagnoses in the latter subgroup (Table S6). Also, positive v (≥1) *alone* triggers TCMR in the Banff classification, whereas v was not part of the 8 most important variables in our TCMR classifier. This is probably because v ≥ 1 was rarely observed in our large training dataset (21/631 cases). Furthermore, among these 21 positive v cases, 10 were not diagnosed as TCMR by the expert group of pathologists and physicians[18] (most of them had $t = 0$ and i = 0, one had $t = 1$). Among the other 11 cases with v ≥ 1 and a reference diagnosis of TCMR, many would have already been considered as borderline TCMR ($n = 2$) or TCMR ($n = 6$) using *only* the i and t scores. In addition, v was missing in a few cases (3.3%) of the training dataset. Borderline TCMR was considered as TCMR in the learning and validation phases of the present study to propose a sensitive tool to detect rejection, considering the clinical "cost" of false negatives higher than that of false positives. Despite a potentially larger variability across centers for reporting borderline TCMR, the accuracy of our TCMR classifier was very high.

TABLE 2 Thresholds chosen for, and performance of, the different classifiers in the Leuven dataset

	ABMR model		TCMR model	
Threshold	Low = 0.10	High = 0.75	Low = 0.10	High = 0.75
Sensitivity (%)	91.7	54.1	91.7	84.8
NPV (%)	99.3	96.8	97.7	96.8
Specificity (%)	97.8	97.9	76.8	97.9
PPV (%)	35.0	64.9	46.1	89.7
Balanced accuracy (%)	94.8	76.0	84.3	91.4
% cases between the two thresholds	11.8		18.5	

Note: For each model, the lower cutoff was chosen to exclude each type of diagnosis with near certainty (to favor sensitivity and the negative predictive value), and the higher cutoff to assert the diagnosis with similar near certainty (to favor specificity and the positive predictive value). Consequently, sensitivity with the higher cutoff and specificity with the lower are only provided (in gray) for completeness.

Abbreviations: ABMR, active antibody-mediated rejection; Balanced accuracy = (Sensitivity + Specificity) / 2; NA, not applicable; NPV, negative predictive value; PPV, positive predictive value; TCMR, T cell–mediated rejection.

In the ABMR model (yes/no), cg was reported as the third most important variable, whereas the Banff classification uses this criterion only for the distinction between active and chronic active ABMR. Moreover, the v criterion was underused (it is not one of the eight most important variables), whereas in the Banff classification when both g and ptc =0, ABMR can be diagnosed based on (for example) v ≥ 1 + C4d ≥ 1 on paraffin sections. In our training dataset, among biopsies with v ≥ 1 ($n = 21$), only two showed both g and ptc scores =0, both negative for C4d and DSA. Logically, these two cases were not retained as ABMR by the expert group. For this reason, the classifier did not consider v for the diagnosis of ABMR, even though it learned on many diverse ABMR cases from real life ($n = 104$ in the training set). In the bigger external dataset from Leuven ($n = 3744$), where diagnoses were made by strictly applying the Banff rules, v ≥ 1 with both g and ptc =0 occurred in 111 biopsies, among which active ABMR was only diagnosed 10 times. The classifier considered 9 out of these 10 cases as *not* being ABMR since time posttransplant, serum creatinine, and proteinuria had higher importance than v. However, to avoid false negatives, a computational layer was added to the TCMR and ABMR classifiers to issue a warning for v-only cases where rejection is not diagnosed, to the effect that the reader may still consider rejection based on the Banff rules.

Regarding the reference ABMR cases without DSA, 88% were diagnosed correctly by the ABMR classifier. However, such cases are not consensually accepted as equivalent to those with positive DSA, since graft survival is not the same.[19] In BIOMARGIN, the expert group of nephrologists carefully labeled these cases as ABMR, sometimes using additional clinical data. The classifier learned from these reference diagnoses and concluded quite accurately about ABMR, based on the features reported in the three large independent datasets (Table 3).

In addition to some major deviations between algorithm predictions and the strict application of the Banff classification presented above, Table 5 shows average differences, from which minor deviations can also be inferred. Based on these findings, several features should perhaps be considered differently for potential improvements to the Banff classification: positive v might not be sufficient on its own for active TCMR, or as a replacement for g and ptc for active ABMR; positive DSA and/or C4d may be given lower weight for ABMR when specific lesions are present; in particular, cg lesions alone may be specific enough to diagnose ABMR even if DSA are negative (Tables 4 and 5); higher serum creatinine and later posttransplant periods may be considered as reinforcing criteria in ambiguous TCMR cases (Table 4).

In the last stage of external validation, our ABMR algorithm was much more sensitive than the Banff classification in detecting ABMR as defined by the MHH expert pathologists, who considered many biopsies as positive despite negative DSA and C4d, based on impaired kidney function and markedly increased histologic criteria specific for ABMR (Table 3). In the Necker database, several TCMR were diagnosed despite missing i or i0, based on t3. Sensitivity with both Banff and our classifier was low in these cases (but better with our algorithm). Of course, such discrepancies may sometimes come from erroneous local diagnoses, suggesting that our algorithms may be useful in influencing the pathologists' calls in these situations. In addition, the threshold selected critically influences the performance of the algorithms toward either sensitivity (by lowering the threshold) or specificity, as shown in Table 3. In the KU Leuven dataset, which was used to set up these thresholds, we observed for both ABMR and TCMR that true negatives were uniformly distributed below 0.10 with a mode very close to 0, whereas the scores of true positive cases were more spread out between 0.75 and 1.

Grading elementary lesions is not always possible, because the number of glomeruli and arteries visible on the slides is sometimes too small. When criteria are missing, classifier performances might be degraded. For routine practice, an overlay of "if then else" rules should be applied upstream of the classifiers to avoid making predictions in case one of these critical determinants, pointed out by the importance plot, is missing. In contrast, even in the absence of one or a few minor predictors (e.g., DSA, time after transplant, C4d staining, serum creatinine, or proteinuria for ABMR), classifier predictions remained accurate.

TABLE 3 Comparison between strict application of the Banff classification and ML algorithms to reproduce the MHH and Necker Hospital experts' diagnoses

	MHH pathologists' expertise as reference ($n = 589$)								Necker hospital pathologists' expertise as reference ($n = 360$)							
	ABMR				TCMR				ABMR				TCMR			
	Banff classification	XGBoost model			Banff classification	XGBoost model			Banff classification	XGBoost model			Banff classification	XGBoost model		
Threshold		0.50	0.10	0.75		0.50	0.10	0.75		0.50	0.10	0.75		0.50	0.10	0.75
Sensitivity (%)	11	97	97	92	86	84	90	82	74	95	99	90	49	60	92	43
Negative predictive value (%)	95	100	100	100	92	91	94	90	92	98	99	97	93	94	98	92
Specificity (%)	100	99	96	100	98	98	92	98	88	82	65	91	95	95	81	92
Positive predictive value (%)	67	83	59	94	95	96	87	97	65	63	47	76	61	64	41	63
Balanced accuracy (%)	55	98	96	96	92	91	91	90	81	89	82	90	72	77	86	67

Note: For each model, the neutral threshold at 0.50 allows making decisions in all cases (no gray zone), the lower cutoff excludes each type of diagnosis with near certainty (to favor sensitivity and the negative predictive value), and the higher cutoff asserts the diagnosis with similar near certainty (to favor specificity and the positive predictive value). Consequently, sensitivity with the higher cutoff and specificity with the lower are only provided (shaded) for completeness.

Abbreviations: ABMR, antibody-mediated rejection; Balanced accuracy = (Sensitivity + Specificity) / 2; TCMR, T cell–mediated rejection; XGBoost, extreme gradient boosting.

TABLE 4 Combinations of features associated with the scores in the "gray" zone

						Average score	Tendencies (based on the median)
ABMR gray scores	positive dsa					0.45 (17%)	In favor of the diagnosis
	negative dsa	cg1+				0.45 (10%)	In favor of the diagnosis
		cg0	g3			0.42 (8%)	In favor of the diagnosis
			g2-	ptc1+	g1+	0.60 (6%)	In favor of the diagnosis
					g0	0.18 (16%)	Against the diagnosis
				ptc0		0.21 (44%)	Against the diagnosis
						Median: 0.24	
						Average score	**Tendencies (based on the median)**
TCMR gray scores	1 year+ post tx	ct1-	S creat 138+ µM	positive dsa		0.66 (1%)	In favor of the diagnosis
				negative dsa	g1+	0.54 (4%)	In favor of the diagnosis
					g0	0.44 (24%)	In favor of the diagnosis
			S creat <138 µM	positive dsa		0.54 (1%)	In favor of the diagnosis
				negative dsa	S creat 218+ µM	0.44 (4%)	In favor of the diagnosis
					S creat <218 µM	0.33 (15%)	In favor of the diagnosis
		ct2+	positive dsa			0.44 (1%)	In favor of the diagnosis
			negative dsa			0.25 (11%)	Against the diagnosis
	< 1 year post tx	t1+				0.68 (1%)	In favor of the diagnosis
		t0	g1+	positive dsa		0.51 (2%)	In favor of the diagnosis
				negative dsa		0.36 (4%)	In favor of the diagnosis
			g0	S creat 212+ µM		0.29 (10%)	Against the diagnosis
				S creat <212 µM		0.20 (21%)	Against the diagnosis
						Median: 0.32	

Note: Cutoff were automatically optimized to distinguish groups of biopsies regarding their scores. Therefore, among all the features available, those most associated with the scores in the "gray" zone were selected to build the trees. At the far right are those with the greatest influence on a given score. The tree complexity was capped, and the tree depth was limited to 5 nodes. Average scores are presented as gradients of blue (step = 0.2).

Abbreviations: ABMR, antibody-mediated rejection; cg, chronic transplant glomerulopathy; ct, tubular atrophy in cortex; dsa, donor-specific antibodies; g, glomerulitis; ptc, peritubular capillaritis; S creat, serum creatinine concentration (µmol/L); t, tubulitis in cortical tubules within non-scarred cortex; TCMR, T cell–mediated rejection; tx, transplantation.

TABLE 5 Comparison of the features for ABMR and TCMR cases diagnosed by strict application of the Banff rules (KU Leuven dataset), with those provided by the algorithms

	ABMR scores >0.75 ($n = 204$)	Strict Banff ABMR ($n = 242$)	TCMR scores >0.75 ($n = 629$)	Strict Banff TCMR ($n = 665$)
g	1.92 (1–3)	1.73 (1–3)	0.52 (0–1)	0.52 (0–1)
ptc	1.41 (1–2)	1.05 (0–2)	0.58 (0–1)	0.55 (0–1)
cg	0.34 (0–0)	0.21 (0–0)	0.11 (0–0)	0.11 (0–0)
DSA positivity	53%	58%	16%	17%
Time post tx, months	17 (0–24)	14 (1–24)	17 (1–24)	16 (0–24)
C4d	1.06 (0–3)	1.46 (0–3)	0.47 (0–1)	0.51 (0–1)
Serum creatinine, µmol/L	268 (152–325)	251 (125–286)	263 (132–305)	273 (133–321)
t	1.06 (0–2)	0.73 (0–1)	1.71 (1–2)	1.49 (1–2)
i	1.17 (0–2)	0.78 (0–1)	1.83 (1–3)	1.58 (1–2)
v	0.39 (0–1)	0.39 (0–1)	0.29 (0–1)	0.43 (0–1)
ct	0.81 (0–1)	0.76 (0–1)	0.86 (0–1)	0.86 (0–1)
ci	0.63 (0–1)	0.60 (0–1)	0.72 (0–1)	0.72 (0–1)
ah	0.73 (0–1)	0.55 (0–1)	0.60 (0–1)	0.62 (0–1)
cv	0.80 (0–1)	0.75 (0–1)	0.73 (0–1)	0.78 (0–1)

Note: Means (interquartile range) are presented here.

Abbreviations: ABMR, antibody-mediated rejection; ah, arteriolar hyalinosis; C4d, linear C4d staining in ptc or medullary vasa recta; cg, chronic transplant glomerulopathy; ci, interstitial fibrosis in cortex; ct, tubular atrophy in cortex; cv, arterial intimal fibrosis (fibrointimal thickening); DSA, donor-specific antibodies; g, glomerulitis; i, inflammation in non-scarred cortex; ptc, peritubular capillaritis; t, tubulitis in cortical tubules within non-scarred cortex; TCMR, T cell–mediated rejection; ti, total cortical inflammation; v, endarteritis (intimal arteritis).

TABLE 6 Evaluation of the XGBoost estimations of active/chronic active ABMR as compared with local diagnosis in the KU Leuven dataset ($n = 232$)

			Local diagnosis	
ABMR predicted in the "gray" zone (0.10 ≤ score <0.75)			**Active ABMR $N = 79$**	**Chronic active ABMR $N = 13$**
Model predictions	Active ABMR	$N = 81$	79	2
	Chronic active ABMR	$N = 11$	0	11
ABMR predicted positive (scores ≥ 0.75)			**Active ABMR $N = 112$**	**Chronic active ABMR $N = 19$**
Model predictions	Active ABMR	$N = 113$	111	2
	Chronic active ABMR	$N = 18$	1	17

Abbreviation: ABMR, antibody-mediated rejection.

Interobserver reproducibility of kidney graft rejection diagnoses was repeatedly assessed, sometimes limited to the detection and grading of elementary lesions (the diagnoses being derived centrally using the Banff rules)[2,4,20] while at other times encompassing the final diagnoses.[21,22] The inter-observer agreement on the conclusions drawn from the semi-quantitative criteria and the clinical context (as is done in routine practice) was not evaluated in any of these studies. Unfortunately, the vulnerability of the Banff classification to misinterpretations has already been demonstrated, especially for antibody-mediated rejection.[5] Our classifiers overcame such inconsistencies, as shown by their full agreement with the 6 case-based scenarios used by Schinstock et al.[5] for a large survey, when the elementary Banff lesions and clinical background were provided. In contrast, the original study showed that case interpretation by 95 clinicians and 72 renal pathologists differed from the reference standards by 26.1% and 35.5%, respectively.

In retrospect, the present study also points out the imperfect reproducibility of case classification within and across large European kidney transplantation centers. It also highlights how AI can support the Banff interpretation of elementary lesions and help pathologists in their routine practice, as well as to minimize outcome uncertainty in multicenter clinical trials in kidney transplantation.

The main limitation of this approach is that it depends on biopsy sampling, preparation, reading, and elementary lesion grading, that is, on human skills and variability. Our classifiers are not able to standardize all these steps, but they improve the *interpretation* of

elementary lesions and related laboratory test results, assuming that they are accurate. When this is not the case the interpretation is not likely to be accurate either. However, the many AI tools being developed for digital image analysis may soon fill part of this gap[23–25] and represent an alternative to time-consuming and non-reproducible visual scoring. For example, Hermsen et al. demonstrated the applicability of convolutional neural networks to automated histologic analysis of biopsy slides.[26]

Online tools strictly applying the Banff rules are certainly useful, but the combination of image segmentation techniques and models trained by experts, encompassing elementary lesions and other data (as in the present study), may also greatly contribute to standardize kidney graft biopsy interpretation, as recently suggested.[27]

Finally, AI may help and standardize the interpretation of complex clinical situations, such as those grouped under the terms "kidney graft rejection." The classifiers described here can be adjusted to future changes in the Banff criteria and diagnostic entities (such as chronic active TCMR as soon as agreement has been reached). Biopsies of the learning dataset will be re-examined by expert pathologists and new classification algorithms trained.

ACKNOWLEDGMENTS

This project was supported by ERACoSysMed-2, the ERA-Net for Systems Medicine in clinical research and medical practice (project ROCKET, JTC2 29).

DISCLOSURE

The authors of this manuscript have conflicts of interest to disclose as described by the *American Journal of Transplantation*. M. Naesens reports being a scientific advisor to, or Editorial Board member, of several journals and Advisor for the European Medicines Agency.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to confidentiality and ethical restrictions.

ORCID

Marc Labriffe https://orcid.org/0000-0001-5840-8904
Jean-Baptiste Woillard https://orcid.org/0000-0003-1695-0695
Wilfried Gwinner https://orcid.org/0000-0003-1703-893X
Jan-Hinrich Braesen https://orcid.org/0000-0002-2863-3067
Dany Anglicheau https://orcid.org/0000-0001-5793-6174
Marion Rabant https://orcid.org/0000-0001-5696-6478
Priyanka Koshy https://orcid.org/0000-0002-2313-5122
Maarten Naesens https://orcid.org/0000-0002-5625-0792
Pierre Marquet https://orcid.org/0000-0001-7698-0760

REFERENCES

1. Loupy A, Haas M, Roufosse C, et al. The Banff 2019 kidney meeting report (I): updates on and clarification of criteria for T cell– and antibody-mediated rejection. *Am J Transplant*. 2020;20(9):2318-2331. doi:10.1111/ajt.15898
2. Marcussen N, Olsen TS, Benediktsson H, Racusen L, Solez K. Reproducibility of the Banff classification of renal allograft pathology. Inter- and intraobserver variation. *Transplantation*. 1995;60(10):1083-1089.
3. Furness PN, Taub N. Convergence of European renal transplant pathology assessment procedures (CERTPAP) project. International variation in the interpretation of renal transplant biopsies: report of the CERTPAP project. *Kidney Int*. 2001;60(5):1998-2012. doi:10.1046/j.1523-1755.2001.00030.x
4. Furness PN, Taub N, Assmann KJM, et al. International variation in histologic grading is large, and persistent feedback does not improve reproducibility. *Am J Surg Pathol*. 2003;27(6):805-810.
5. Schinstock CA, Sapir-Pichhadze R, Naesens M, et al. Banff survey on antibody mediated rejection clinical practices in kidney transplantation: diagnostic misinterpretation has potential therapeutic implications. *Am J Transplant*. 2019;19(1):123-131. doi:10.1111/ajt.14979
6. Vaulet T, Divard G, Thaunat O, et al. Data-driven derivation and validation of novel phenotypes for acute kidney transplant rejection using semi-supervised clustering. *J Am Soc Nephrol JASN*. 2021;32(5):1084-1096. doi:10.1681/ASN.2020101418
7. Marx D, Metzger J, Olagne J, et al. Proteomics in kidney allograft transplantation-application of molecular pathway analysis for kidney allograft disease phenotypic biomarker selection. *Proteomics Clin Appl*. 2019;13(2):e1800091. doi:10.1002/prca.201800091
8. Rabant M, Amrouche L, Lebreton X, et al. Urinary C-X-C motif chemokine 10 independently improves the noninvasive diagnosis of antibody-mediated kidney allograft rejection. *J Am Soc Nephrol JASN*. 2015;26(11):2840-2851. doi:10.1681/ASN.2014080797
9. Haas M, Sis B, Racusen LC, et al. Banff 2013 meeting report: inclusion of c4d-negative antibody-mediated rejection and antibody-associated arterial lesions. *Am J Transplant*. 2014;14(2):272-283. doi:10.1111/ajt.12590
10. Haas M. The revised (2013) Banff classification for antibody-mediated rejection of renal allografts: update, difficulties, and future considerations. *Am J Transplant*. 2016;16(5):1352-1357. doi:10.1111/ajt.13661
11. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. ArXiv160302754 cs. Published online June 10, 2016. 10.1145/2939672.2939785
12. XGBoost. Accessed April 21, 2021. https://kaggle.com/dansbecker/xgboost
13. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. ICML '06. Association for Computing Machinery; 2006:233–240. 10.1145/1143844.1143874
14. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
15. Greenwell B. *Fastshap: Fast Approximate Shapley Values*. 2020. Accessed October 14, 2021. https://CRAN.R-project.org/package=fastshap
16. Cannesson M. The "grey zone" or how to avoid the binary constraint of decision-making. *Can J Anaesth*. 2015;62(11):1139-1142. doi:10.1007/s12630-015-0465-1
17. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. 2nd ed. Springer-Verlag; 2009. doi:10.1007/978-0-387-84858-7
18. Wohlfahrtova M, Hruba P, Klema J, et al. Early isolated V-lesion may not truly represent rejection of the kidney allograft. *Clin Sci*. 2018;132(20):2269-2284. doi:10.1042/CS20180745
19. Senev A, Coemans M, Lerut E, et al. Histological picture of antibody-mediated rejection without donor-specific anti-HLA antibodies: clinical presentation and implications for outcome. *Am J Transplant*. 2019;19(3):763-780. doi:10.1111/ajt.15074

20. Smith B, Cornell LD, Smith M, et al. A method to reduce variability in scoring anti-body mediated rejection in renal allografts: implications for clinical trials. *Transpl Int*. 2019;32(2):173-183. doi:10.1111/tri.13340
21. Gough J, Rush D, Jeffery J, et al. Reproducibility of the Banff schema in reporting protocol biopsies of stable renal allografts. *Nephrol Dial Transplant*. 2002;17(6):1081-1084. doi:10.1093/ndt/17.6.1081
22. Veronese FV, Manfro RC, Roman FR, et al. Reproducibility of the Banff classification in subclinical kidney transplant rejection. *Clin Transplant*. 2005;19(4):518-521. doi:10.1111/j.1399-0012.2005.00377.x
23. Gadermayr M, Dombrowski AK, Klinkhammer BM, Boor P, Merhof D. CNN cascades for segmenting sparse objects in gigapixel whole slide images. *Comput Med Imaging Graph*. 2019;71:40-48. doi:10.1016/j.compmedimag.2018.11.002
24. Pedraza A, Gallego J, Lopez S, Gonzalez L, Laurinavicius A, Bueno G. Glomerulus classification with convolutional neural networks. In: Valdés Hernández M, González-Castro V, eds. *Medical Image Understanding and Analysis*. Communications in Computer and Information Science. Springer International Publishing; 2017:839–849. 10.1007/978-3-319-60964-5_73
25. Bukowy JD, Dayton A, Cloutier D, et al. Region-based convolutional neural nets for localization of glomeruli in trichrome-stained whole kidney sections. *J Am Soc Nephrol*. 2018;29(8):2081-2088. doi:10.1681/ASN.2017111210
26. Hermsen M, de Bel T, den Boer M, et al. Deep learning-based histopathologic assessment of kidney tissue. *J Am Soc Nephrol*. 2019;30(10):1968-1979. doi:10.1681/ASN.2019020144
27. Arthurs C, Roufosse C. Forging the tools for a computer-aided workflow in transplant pathology. *Lancet Digit Health*. 2022;4(1):e2-e3. doi:10.1016/S2589-7500(21)00254-5

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Labriffe M, Woillard J-B, Gwinner W, et al. Machine learning-supported interpretation of kidney graft elementary lesions in combination with clinical data. *Am J Transplant*. 2022;00:1-13. doi:10.1111/ajt.17192

SUPPORTING INFORMATION

Patients and biopsies

For the training set, we used biopsy data from two European programs, BIOMArkers of Renal Graft INjuries (BIOMARGIN, ClinicalTrials.gov, number NCT02832661) and Reclassification using OmiCs integration in KidnEy Transplantation (ROCKET, funded by ERACoSysMed 2018-2021), both aiming at discovering and validating robust non-invasive biomarkers. The first two steps of BIOMARGIN were case-control studies enabling an untargeted search and then the selection of a broad list of biomarkers. The third, cross-sectional step, aimed to validate the diagnostic performance of the biomarker candidates on a representative sample of transplant patients in Europe. Between June 2011 and August 2016, more than 650 sample triplets (urine, blood and biopsy) were collected in highly standardized conditions and stored in the Biobanks of the four hospitals participating in the project (Hôpital Necker Paris, France; University Hospitals Leuven, Belgium; Medizinische Hochschule Hannover, Germany; and Centre Hospitalier Universitaire Limoges, France).

Biopsy and omics data are still being gathered in our consortium as part of the ROCKET program, to discover accurate biomarkers of rarer phenotypes or graft lesions, including: active antibody-mediated rejection (ABMR), chronic active ABMR, acute T cell-mediated rejection (TCMR), chronic active TCMR, polyomavirus nephropathy (PVN) and glomerulonephritis. The ABMR cases of the corresponding dataset were used to train an additional model able to distinguish active from chronic active ABMR. We could not study chronic inactive ABMR because patient's history was not available in these case-control or cross-sectional studies.

For the external validation of the ML classifiers and the choice of thresholds, we first used biopsy data from patients transplanted between 2004 and 2013 and followed-up until September 2019 at KU Leuven, Belgium. The second validation dataset was from patients followed-up from 2013 to 2019 at the Medizinische Hochschule Hannover, Germany and the third from a single-center study at Hôpital Necker, Paris, France, approved by the ethics committee of Ile-de-France XI (13016), where clinically-indicated renal allograft biopsies were collected from February 2011 to February 2013.

Imputation in the training dataset

In the training dataset, biopsies with more than 2 missing data among the elementary Banff lesions were removed. This exclusion was not applied to the validation datasets, in order to evaluate the classifiers in real-life situations. After analyzing the distribution of the Banff elementary lesion scores, we chose to impute the respective median value to the missing scores (when less than 2 per biopsy), in the training dataset. No imputation was made in the different validation datasets.

Definitions of the phenotypes in the different external validation cohorts

University Hospitals Leuven, Belgium

In the Leuven cohort, a non-standard immunohistochemistry stain was used on frozen sections for C4d. Therefore, the positivity threshold was 2/3 for C4d scores, as mentioned in the Banff classification. For the rest of the criteria, the 2019 Banff rules were strictly applied retrospectively to all the biopsies. For example, isolated v was considered as TCMR grade II, which is still under debate.[1] Donor specific antibodies (DSA) detection was performed using the Immucor kit on Luminex.

Medizinische Hochschule Hannover, Germany

In the Hannover cohort, the TCMR diagnosis was corrected when sv40 staining was positive (to exclude false positives due to BK virus nephropathy -BKVN-). Indeed, in practice, when BK viremia is positive, pathologists systematically test for BKVN using SV40 staining.[2] In this cohort, all the BKVN cases had positive BK viremia. DSA detection was also performed using the Immucor kit on Luminex. Mostly, multiple antigen testing was used for screening, single antigen testing afterwards in positive cases. During certain time periods, both tests were run in parallel.

University Hospital Necker Paris, France

In the Necker Cohort, the most recent Banff rules at the time of each biopsy were strictly applied in the majority of the cases, except for adjustments made by the clinicians in complex situations. In this center too, SV40 staining was performed in case of concomitant BK viremia. DSA were detected using the One Lambda Single-antigen Bead Flow Cytometry Assays on Luminex.

Supplemental Table 1: XGBoost hyperparameters.

	Model			
Hyperparameters	Active ABMR	TCMR	IFTA	ABMR active/chronic active
mtry	2	20	15	13
Trees	1000	1000	1000	1000
min_n	2	6	4	2
Tree depth	3	9	4	2
Learning rate	0.024	0.015	0.014	0.017

Abbreviations: Learning rate, rate at which the boosting algorithm adapts from iteration-to-iteration; min_n, minimum number of data points in a node that are required for the node to be split further; mtry, number of predictors that will be randomly sampled at each split when creating the tree models; Tree depth, maximum depth of the tree (i.e., number of splits); Trees, number of trees contained in the boosted ensemble (i.e., number of boosting iterations).

Supplemental Table 2: Diagnostic characteristics for the cases used in the training step.

Variables	Normal (BIOMARGIN) N = 312	ABMR (BIOMARGIN) N = 104	TCMR (BIOMARGIN) N = 82	IFTA (BIOMARGIN) N = 210	Missing data (BIOMARGIN) N = 631 (100%)	Active ABMR (ROCKET) N = 63	Chronic active ABMR (ROCKET) N = 44
g, mean (IQR)	0.03 (0-0)	1.52 (1-2)	0.44 (0-1)	0.28 (0-0)	0 (0.0%)	1.38 (1-2)	1.50 (1-2)
ptc, mean (IQR)	0.05 (0-0)	1.53 (1-2)	0.78 (0-1)	0.40 (0-0)	1 (0.2%)	1.75 (1-2)	0.87 (1-2)
v, mean (IQR)	0.00 (0-0)	0.19 (0-0)	0.23 (0-0)	0.03 (0-0)	21 (3.3%)	0.32 (0-0)	0.09 (0-0)
C4d, mean (IQR)	0.22 (0-0)	0.79 (0-2)	0.39 (0-0)	0.25 (0-0)	21 (3.3%)	0.87 (0-2)	0.59 (0-1)
DSA positivity (%)	39 (13.2)	56 (54.9)	30 (37)	33 (16.4)	27 (4.3%)	35 (55.6)	21 (47.7)
cg, mean (IQR)	0.01 (0-0)	1.01 (0-2)	0.24 (0-0)	0.29 (0-0)	6 (1.0%)	0.08 (0-0)	2.18 (2-3)
i, mean (IQR)	0.02 (0-0)	0.38 (0-0)	1.29 (0-2)	0.20 (0-0)	1 (0.2%)	0.57 (0-1)	0.27 (0-0)
t, mean (IQR)	0.06 (0-0)	0.42 (0-1)	1.49 (1-2)	0.27 (0-0)	1 (0.2%)	0.62 (0-1)	0.27 (0-1)
ti, mean (IQR)	0.05 (0-0)	0.66 (0-1)	1.48 (1-3)	0.79 (0-1)	9 (1.4%)	0.70 (0-1)	0.66 (0-1)
ct, mean (IQR)	0.32 (0-1)	1.17 (0-2)	1.15 (0-2)	2.43 (2-3)	1 (0.2%)	0.81 (0-1)	1.64 (1-3)
ci, mean (IQR)	0.32 (0-1)	1.16 (0-2)	1.07 (0-2)	2.40 (2-3)	1 (0.2%)	0.84 (0-2)	1.59 (1-3)
ah, mean (IQR)	0.47 (0-1)	1.31 (0-3)	0.80 (0-1)	1.37 (0-2)	9 (1.4%)	0.78 (0-1)	1.98 (1-3)
cv, mean (IQR)	0.70 (0-1)	1.33 (0-3)	1.09 (0-2)	1.58 (1-2)	33 (5.2%)	0.89 (0-2)	1.80 (1-3)
Serum creatinine (µmol/L), mean (IQR)	154 (110-169)	222 (149-240)	226 (131-254)	192 (133-229)	0 (0.0%)	225 (127-237)	217 (156-254)
Proteinuria (g/L), mean (IQR)	0.20 (0.07-0.15)	0.87 (0.07-1.13)	0.50 (0.07-0.32)	0.40 (0.07-0.30)	0 (0.0%)	0.41 (0.07-0.51)	1.49 (0.25-2.29)
Time after transplant (mo), mean (IQR)	17 (3-12)	70 (5-120)	35 (3-53)	50 (12-60)	0 (0.0%)	20 (3-25)	129 (47-186)

Abbreviations: ABMR, active antibody-mediated rejection; ah, arteriolar hyalinosis criterion; c4d, linear C4d staining in ptc or medullary vasa recta criterion; cg, chronic transplant glomerulopathy criterion; ci, interstitial fibrosis in cortex criterion; ct, tubular atrophy in cortex criterion; cv, arterial intimal fibrosis criterion (fibrointimal thickening); DSA, donor-specific antibodies; g, glomerulitis criterion; i, inflammation in non-scarred cortex criterion; IFTA, interstitial fibrosis/tubular atrophy grade II; IQR, interquartile range; Normal, refers to cases with no graft alterations; ptc, peritubular capillaritis criterion; t, tubulitis in cortical tubules within non-scarred cortex criterion; TCMR, T cell-mediated rejection; ti, total cortical inflammation criterion; v, endarteritis (intimal arteritis).

Supplemental Table 3: Performance of the IFTA classifier.

	Leuven database	MHH database	Necker database
	Threshold = 0.10		
Sensitivity (%)	100	95.5	99.5
NPV (%)	100	99.6	99.4
Specificity (%)	100	100	94.8
PPV (%)	100	100	95.4
Balanced accuracy (%)	100	97.8	97.2

Abbreviations: Balanced accuracy = (Sensitivity + Specificity) / 2; IFTA, interstitial fibrosis/tubular atrophy grade II; NPV, negative predictive value; PPV, positive predictive value.

Supplemental Table 4: Machine-learning analysis of the 6 case-based scenarios used by Schinstock et al. for their international survey among clinicians and renal pathologists.

Case-based scenarios																		Machine-learning interpretation							
Case #	g	ptc	c4d	cg	v	DSA	i	t	ti	ct	ci	ah	cv	Serum creatinine (µmol/L)	Proteinuria	Time between transplantation and biopsy (months)	Reference diagnosis	ABMR score	Active ABMR prediction	Chronic active ABMR score	Chronic active ABMR prediction	TCMR score	TCMR prediction	IFTA score	IFTA prediction
1	2	1	1	1	MD	1	0	0	MD	0	0	MD	MD	194.52	MD	48	Chronic active ABMR	0.99	positive	0.84	positive	0.01	negative	0.00	negative
2	1	2	0	MD	MD	1	0	0	MD	0	MD	MD	MD	123.79	MD	36	Acute/active ABMR	0.98	positive	0.04	negative	0.02	negative	0.01	negative
3	1	2	0	1	MD	1	0	0	MD	0	0	MD	MD	194.52	MD	120	Chronic active ABMR	1.00	positive	0.85	positive	0.01	negative	0.00	negative
4	2	2	0	MD	MD	0	0	0	MD	0	0	MD	MD	150.31	MD	12	Histologic features of ABMR without detectable anti-HLA antibody	0.96	positive	0.05	negative	0.02	negative	0.01	negative
5	2	2	0	MD	MD	1	0	0	MD	0	0	MD	MD	150.31	MD	6	Acute/active ABMR	0.99	positive	0.05	negative	0.01	negative	0.01	negative
6	1	2	0	0	MD	1	2	3	MD	MD	MD	MD	MD	221.05	MD	18	Mixed acute T cell mediated rejection and ABMR	0.98	positive	0.01	negative	0.98	positive	0.00	negative

Abbreviations: ah, arteriolar hyalinosis criterion; c4d, linear C4d staining in ptc or medullary vasa recta criterion; cg, chronic transplant glomerulopathy criterion; ci, interstitial fibrosis in cortex criterion; ct, tubular atrophy in cortex criterion; cv, arterial intimal fibrosis criterion (fibrointimal thickening); dsa, donor-specific antibodies; g, glomerulitis criterion; i, inflammation in non-scarred cortex criterion; MD, missing data; ptc, peritubular capillaritis criterion; t, tubulitis in cortical tubules within non-scarred cortex criterion; ti, total cortical inflammation criterion.

Supplemental Table 5: Decision tree to distinguish borderline from TCMR cases using clinical information, when i and/or t scoring is missing (Banff classification not applicable).

Cutoffs were automatically optimized to distinguish the Borderline and TCMR subgroups, among all available features excluding i, t and v (which is still under debate). Tree depth was capped at 4 nodes. The percentages of TCMR in the final leaves were colored with blue gradients (every 20%).

					TCMR cases	Tendencies
Borderline and TCMR cases	< 3 months post-tx				81%	TCMR
	≥ 3 months post-tx	S creat ≥314 µM			82%	TCMR
		S creat < 314 µM	< 7 years post-tx	cv < 1	60%	TCMR
				cv ≥ 2	43%	Borderline
			≥ 7 years post-tx		0%	Borderline

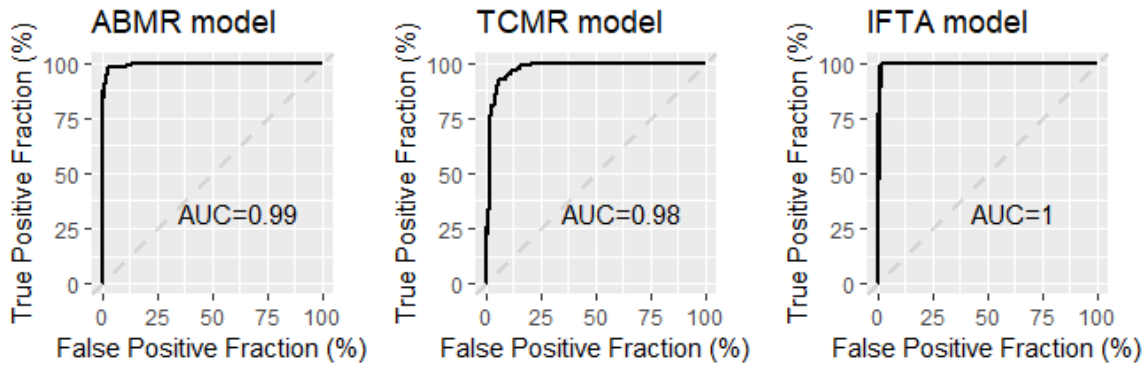
Abbreviations: cv, arterial intimal fibrosis criterion (fibrointimal thickening); S creat, serum creatinine concentration (µmol/L); TCMR, T cell-mediated rejection; tx, transplantation.

Supplemental Table 6: The inflammation in non-scarred cortex criterion (i criterion) and the tubulitis in cortical tubules within non-scarred cortex criterion (t criterion) in the training set and the corresponding TCMR diagnoses of the experts.

Number of cases for each combination of i and t, and percentage of them diagnosed as TCMR by the experts. It is important to note that these diagnoses were made in light of other histologic features and the clinical context.

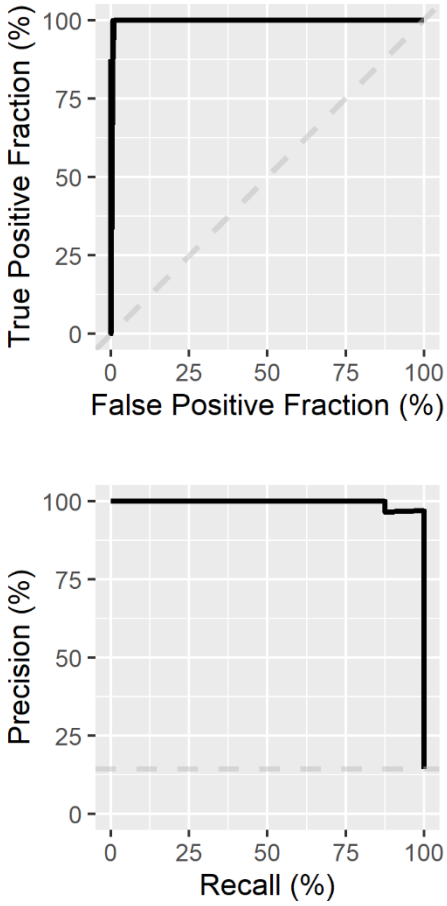
		t				
		0	1	2	3	Total
i	0	n = 511 1% TCMR	n = 36 28% TCMR	n = 5 80% TCMR	n = 4 75% TCMR	n = 556 4% TCMR
	1	n = 6 17% TCMR	n = 27 89% TCMR	n = 6 67% TCMR	n = 4 75% TCMR	n = 43 74% TCMR
	2	n = 0	n = 7 100% TCMR	n = 4 100% TCMR	n = 2 100% TCMR	n = 13 100% TCMR
	3	n = 0	n = 6 83% TCMR	n = 6 83% TCMR	n = 7 86% TCMR	n = 19 84% TCMR
	Total	n = 517 1% TCMR	n = 76 61% TCMR	n = 21 81% TCMR	n = 17 82% TCMR	n = 631 13% TCMR

Abbreviations: i, inflammation in non-scarred cortex criterion; TCMR, T cell-mediated rejection; t, tubulitis in cortical tubules within non-scarred cortex criterion.



Supplemental Figure 1: ROC curve analysis in the training dataset.

Information provided in this figure was obtained using the hyperparameters tuned using the same dataset. After tuning the hyperparameters, the ROC curves showed AUC = 0.99 (95% CI: 0.99-1.00), 0.98 (95% CI: 0.96-0.99) and 1.00 (95% CI: 0.99-1.00) for ABMR, TCMR and IFTA, respectively. The accuracy was 0.97, 0.95, 0.99 and 0.94 for the ABMR model, the TCMR model, the IFTA model and the ABMR active/chronic model, respectively (with thresholds arbitrarily set at 0.50).



Supplemental Figure 2: ROC curve and PR curve analyses for the diagnosis of chronic active ABMR, among the cases with ABMR scores > 0.10 in the Leuven dataset.

The AUC ROC and the PR ROC was 1.00 (with threshold arbitrarily set at 0.50).

References

1. Wohlfahrtova M, Hruba P, Klema J, Novotny M, Krejcik Z, Stranecky V, et al.: Early isolated V-lesion may not truly represent rejection of the kidney allograft. Clin. Sci. Lond. Engl. 1979 132: 2269–2284, 2018

2. Hirsch HH, Brennan DC, Drachenberg CB, Ginevri F, Gordon J, Limaye AP, et al.: Polyomavirus-associated nephropathy in renal transplantation: interdisciplinary analyses and recommendations. Transplantation 79: 1277–1286, 2005

IV. Valider un score de risque individuel de base pour la perte du greffon

IV.1. Objectifs

De nombreux outils statistiques ont été développés pour tenter de classer les patients transplantés en fonction de leur pronostic. Quel que soit la méthode statistique utilisée, la plupart des scores existants n'ont pas été testés sur des cohortes longitudinales avec une grande profondeur temporelle. Beaucoup d'articles ont préféré mettre en avant de bons résultats à court-moyen terme, en prédisant des évènements survenant quelques années après la consultation de suivi (plus faciles à prévoir). De plus, ces scores requièrent souvent de nombreuses informations sur le patient, parfois très détaillées comme les scores histologiques des comptes-rendus de biopsie de greffon.

L'objectif principal de ce travail était de valider le score pronostique ajustable prédicteur de l'échec de greffe (AdGFS[12]) préalablement développé par le Pr. Annick Rousseau et le Dr. Aurélie Prémaud dans notre unité Inserm 1248 [50] : simple d'utilisation, validé à court et long terme de manière transparente, sur deux centres européens indépendants, Leuven (n = 930) et Lyon (n = 307).

Les variables sélectionnées par ML, plus précisément par forêts aléatoires de survie, incluaient : l'âge du donneur ; les anticorps anti-HLA non spécifiques du donneur avant la greffe (présence/absence) ; le profil cinétique de la créatinine sérique en utilisant les valeurs aux 1er, 3^{e}, 6^{e} et 12^{e} mois post-transplantation ; la créatinine sérique et la protéinurie (g/L ou g/24h) à un an post-transplantation ; l'apparition d'anticorps anti-HLA spécifiques du donneur (oui/non) ; la survenue d'un premier épisode de rejet aigu. Ces variables avaient ensuite servi à construire un arbre de décision. En fonction des branches, 2 à 7 variables étaient utilisées pour obtenir un score pronostic coté de 0 à 12 (0 étant le meilleur pronostic).

IV.2. Discussion

Dans cet Article 4, nous avons validé un score pronostique, présenté sous la forme d'un arbre de décision distribuant des points : AdGFS. Le suivi médian des patients était long : jusqu'à 9 ans, ce qui a permis d'évaluer les performances de cet outil avec une grande profondeur temporelle. L'évaluation était graphique, par les courbes de Kaplan-Meier, montrant une stratification des profils de risques en fonction des valeurs prises par le score. Le score était également évalué sur les AUC des courbes ROC, qui par définition prennent en compte plusieurs seuils, et qui étaient représentées en fonction du temps. Enfin, les valeurs de sensibilité, spécificité, valeur prédictive positive, et valeur prédictive négative ont été détaillées dans le Tableau 2 de l'article, avec plusieurs seuils et à différents temps post-transplantation.

[12] AdGFS : adjustable score for prediction of graft failure.

Dans une publication récente, Truchot et al. n'ont pas mis en évidence d'amélioration significative de performance entre des algorithmes complexes de ML et des modèles de survie classiques (à 7 ans post transplantation) [141]. L'objectif n'était probablement pas de critiquer l'utilisation de l'IA puisque certains des auteurs développent dans le même temps des outils pour recueillir plus de données, en défendant explicitement une future utilisation de l'IA pour réaliser des prédictions [142].

Dans leur article, le temps de suivi médian était plus court que dans notre étude : 6 ans versus 9 ans. Dans notre cas, le ML (forêt aléatoire de survie) était employé uniquement pour faire de la sélection de variables. Truchot et al. ont entraîné des modèles complexes : forêts aléatoires de survie (RSF), machines à vecteurs de support (SVM), et XGBoost. Ils ont utilisé de *nombreuses* données du donneur, du receveur, et du greffon. Avec le modèle de Prémaud et al. [50], l'objectif était différent : un *simple* arbre de décision de survie, utilisable facilement en pratique, avec des données toujours et partout disponibles. Au final, avec un modèle plus simple à manipuler, et moins de prédicteurs à renseigner, la performance prédictive était proche de 0,75 pour notre modèle, versus 0,81 pour leur meilleur modèle (à 7 ans post transplantation).

Le système DISPO[13] de Raynaud et al. [143] nécessite lui aussi plus de variables que AdGFS : au moins 11, en plus des valeurs de suivi du DFG estimé ou de la protéinurie. Ces variables comprennent 5 critères de Banff qui ne sont disponibles que si des biopsies ont été réalisées. DISPO a été développé et validé à l'aide de plusieurs bases de données cliniques de patients transplantés rénaux suivis pendant une moyenne de 7 (4-10) ans. Malgré cela, l'application en ligne fournit des probabilités de survie jusqu'à 9 ans, avec 3 chiffres décimaux et sans intervalle de confiance [144]. Nous avons comparé les performances des deux outils en utilisant les informations de la première année de suivi après la biopsie (DISPO) ou de la première année post-transplantation (AdGFS), dans la population européenne, pour prédire l'événement 5 ans plus tard et avons trouvé des AUC ROC = 0,80 et 0,72, respectivement. Lorsque l'événement se produisait 6 ans plus tard, l'AUC était de 0,76 avec DISPO. Les performances à plus long terme n'ont pas été étudiées. Les résultats des figures de l'étude de Raynaud et al. concernaient des prédictions faites à partir de données accumulées sur 2, 3, 4, 5 ans, pour prédire un événement plus proche dans le temps : 4, 3, 2, 1 ans plus tard, respectivement. Au final, le résultat principal présenté par les auteurs de cet article est la moyenne de ces 5 valeurs ROC AUC (comme expliqué dans l'annexe de leur article), gonflant le résultat de manière favorable.

Loupy et al. [49] ont proposé un autre score appelé iBox. Malheureusement, nous n'avons pas pu le tester car le lien du site Web est corrompu [49]. À l'instar de DISPO, le score iBox nécessite 10 variables, y compris les informations de la première biopsie. La plupart des premières biopsies ont eu lieu après la première année post-transplantation (jusqu'à 8 ans, comme présenté dans l'annexe de leur article). Il était donc plus facile de faire des prévisions à long terme puisque des données primordiales étaient recueillies à distance de la transplantation.

[13] DISPO : dynamic, integrative system for predicting outcome.

Au final, le score AdGFS permet d'*anticiper* le risque de perte du greffon à long terme, à partir de caractéristiques recueillies assez tôt pendant le suivi (au bout de la première année post transplantation). Ce score peut servir de profil de risque *de référence* pour orienter les différentes stratégies thérapeutiques qui peuvent être proposées au patient transplanté. Par ailleurs, il est évolutif et peut-être recalculé régulièrement et/ou à l'occasion d'évènements indésirables (rejets, comorbidités).

Nous travaillons actuellement à la construction d'un outil pronostic plus élaboré, grâce au pipeline AutoPrognosis [145]. Nous utiliserons une base de données de patients ayant bénéficié d'un suivi plus long, et les informations des biopsies réalisées pendant la 1re année post-transplantation seront aussi prises en compte.

IV.3. Article 4

Validation of the adjustable score for prediction of kidney graft failure in two European retrospective cohorts: an easy-to-use tool for short to long-term predictions

Manuscrit soumis pour publication à PLOS One.

External validation of an easy-to-use, adjustable score for short to long-term prediction of kidney graft failure.

Marc Labriffe[1,2], Aurélie Prémaud[1], Maarten Naesens[3,4], Olivier Thaunat[5,6], Jean-Baptiste Woillard[1,2], and Pierre Marquet[1,2*]

[1]Inserm, Univ. Limoges, CHU Limoges, Pharmacology & Transplantation, U 1248, F-87000 Limoges, France

[2]CHU Limoges, Service de Pharmacologie Toxicologie et Pharmacovigilance, F-87000 Limoges, France

[3]Nephrology and Renal Transplantation Research Group, Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium

[4]Department of Nephrology and Renal Transplantation, University Hospitals Leuven, Leuven, Belgium

[5]International Center of Infectiology research (CIRI), French Institute of Health and Medical Research (INSERM) Unit 1111, Claude Bernard University Lyon I, National Center for Scientific Research (CNRS) Mixed University Unit (UMR) 5308, Ecole Normale Supérieure de Lyon, University of Lyon, Lyon, France

[6]Department of Transplantation, Nephrology and Clinical Immunology, Hospices Civils de Lyon, Edouard Herriot Hospital, Lyon, France

* **Corresponding Author**

Prof. Pierre Marquet, Department of Pharmacology, Toxicology and Pharmacovigilance

University Hospital of Limoges, CBRS, 2 rue Bernard Descottes, 87000 Limoges, France

Email: pierre.marquet@unilim.fr; Tel: +33 555 05 64 18; ORCID: 0000-0001-7698-0760

ORCID of the authors

Marc Labriffe: 0000-0001-5840-8904

Aurélie Prémaud: 0000-0001-5004-5918

Maarten Naesens: 0000-0002-5625-0792

Olivier Thaunat: 0000-0002-3648-8963

Jean-Baptiste Woillard: 0000-0003-1695-0695

Pierre Marquet: 0000-0001-7698-0760

Conflicts of Interest statement

The authors have no conflicts of interest to declare.

Short title

Validation of the adjustable kidney graft failure score

Keywords

graft survival; renal transplantation; score; creatinine

Abstract

Many tools have been developed to estimate the risk of graft loss for kidney transplant patients, but only a few have been thoroughly validated. We previously developed the adjustable graft failure score (AdGFS), an easy-to-use tool combining 2-7 variables routinely collected at one-year post-transplant in all of the transplantation centers. During model development, we selected the variables using random survival forest, a Machine Learning method, to come up with a simple decision tree based on: donor age; pretransplant non-donor-specific anti-HLA antibodies (presence/absence); kinetic profile of serum creatinine using its values at months 1, 3, 6 and 12 post-transplant; serum creatinine and proteinuria (g/L or g/24h) at one year post-transplantation; occurrence of de novo donor-specific anti-HLA antibodies (yes/no); occurrence of a first episode of acute rejection. The aim of the present study was to validate AdGFS on two large, independent, longitudinal cohorts of kidney transplant patients, from Leuven, Belgium (n = 930 patients) and Lyon, France (n = 307 patients). The 4 risk groups obtained actually showed very large and significant differences in graft survival at 10 years post-transplantation: 89 %, 81 %, 59 % and 23 % in the low, intermediate, high and very high-risk groups, respectively. At 2, 5 and 10 years post-transplantation, the AUC (95% CI) of the ROC curve of AdGFS was 0.86 (0.76-0.96), 0.76 (0.70-0.82), and 0.73 (0.69-0.77), respectively. This external validation study confirmed that AdGFS is clinically applicable for the long-term risk graduation of graft loss at the first transplantation anniversary. It suggests that it may be used to tailor the immunosuppressive regime to the individual risk, and as a surrogate outcome in clinical trials.

Introduction

In patients with terminal renal failure, kidney transplantation greatly improves quality of life [1] and survival [2], as compared with long-term dialysis treatment. However, the success of kidney transplantation is heterogeneous, and the therapeutic management of certain patients requires special attention. Patient outcomes depend on many risk factors. Some of them are baseline characteristics, available before or shortly after transplantation: donor age [3], the presence of pretransplant donor-specific (DSA) or non-donor-specific (NDSA) anti-HLA antibodies [4], graft function as evaluated through serum creatinine (Scr) [5] as well as proteinuria [6] during the first months post-transplantation. Other risk factors may occur at any time after transplantation and modify patient prognosis: donor-specific antibodies (DSA) [7,8], episodes of acute graft rejection, graft infection, comorbidities, etc. [9]. All must be taken into consideration when forecasting graft survival (or graft function deterioration) and patient survival.

In a previous work, we developed a score capable of stratifying, at one year post-transplant, patients into four risk-groups of graft survival over the following 10 years [10]. It is called the adjustable score for prediction of graft failure (AdGFS) since the patient score can evolve with time depending on medical events. It was developed in a cohort of 664 patients transplanted at Limoges University Hospital and externally validated on an external cohort of 896 patients from Poitiers and Tours University Hospitals, France.

Although a myriad of prediction tools have been and are probably still being developed for medical applications, only a small number have been externally validated in the field of nephrology [11]. Such validation is highly sought after for routine clinical use. In the present

study, we aimed to validate AdGFS on two additional and more recent, independent cohorts of kidney transplant patients.

Methods

Score

The scoring system for computing AdGFS is simple and has already been summarized in a figure in a previous paper (Supporting information) [10]. It uses the following variables: donor age (years); pretransplant NDSA (presence/absence); kinetic profile of Scr (µM) during the first year post-transplantation drawn and stratified using its values at months 1 (M1), 3 (M3), 6 (M6) and 12 (M12) post-transplant; Scr and proteinuria (g/L or g/24h) at M12; occurrence of DSA (yes/no) or occurrence of a first episode of acute rejection (antibody-mediated rejection or T-Cell mediated rejection / absence) over the first year post-transplant. During model development, these variables were selected using random survival forest, a Machine Learning method. At each node, depending on the patient status, points are allocated (+0, +2, +4, +10) to produce a score ranging from 0 to 12. Four risk groups were defined according to the AdGFS value: low risk (0), intermediate risk (2-4), high risk (6-8), and very high risk (10-12).

Patients

In this study, we analyzed databases from two kidney transplant centers: KU Leuven (Belgium) and HCL Lyon (France). Patients were transplanted from March 2004 to February 2013 (n = 930) in Leuven and from January 2005 to December 2012 (n = 307) in Lyon and were followed-up for a minimum of 12 months.

Data were de-identified and the collection of clinical data and laboratory test results was declared to the French Ministry of Health (N° AC-2016-2758).

Database Description

During the development of the model, 3 clusters of Scr time-evolution were created based on k-means of all the values available (clusters A, B and C) [10]. With this statistical method, a *new* profile can be classified in one of the three predefined clusters. It only requires the clusters' centers, provided in Supporting information. For each transplant (since a few patients were transplanted twice in the time period considered), all Scr at M1 ± 10 days, M3 ± 2 weeks, M6 ± 3 weeks and M12 ± 4 weeks post-transplantation were extracted to allocate the right cluster. Missing values were replaced by the mean of the previous and the next Scr (e.g., missing M3 Scr was imputed as the mean of M1 and M6 Scr). When M12 proteinuria in g/L was missing, we used the M12 proteinuria in g/24h, as described by Prémaud et al. [10].

Statistical Analyses

For all survival analyses, data were censored at the time of the first event of interest, either return to dialysis and death (all causes) with a non-functioning graft. Survival curves were plotted using the Kaplan-Meier method. Time-dependent sensitivity and specificity for censored event-times were estimated using the definition of Heagerty and Zheng [12], and explained by Kamarudin et al. [13], for cumulative sensitivity and dynamic specificity. Accuracy was calculated with the usual formula:

$$Accuracy = \frac{true\ positives + true\ negatives}{all\ patients}$$

For statistical computing and graphics, we used R (version 4.0.3), in particular the kml package (version 2.4.1) for cluster attribution, the timeROC package (version 0.4) for estimation of the areas under the time-dependent receiver operating characteristic curve (ROC AUC) and its confidence interval (CI), and the survival package (version 3.2-7) for survival analyzes.

Results

The characteristics of the studied population are presented in Table 1. The median duration of follow-up was 9 years for the 1237 patients of the study, with a total of 222 events: return to dialysis or death with a non-functioning graft.

Table 1. Kidney transplant characteristics of the Leuven and Lyon databases

Medians [interquartile ranges] are presented here.

DSA, donor-specific anti-HLA antibodies; M12, month 12 post-transplantation; MD, missing data; NDSA, non-donor-specific anti-HLA antibodies.

	Leuven database (n = 930)	Lyon database (n = 307)	All (n = 1237)
Duration of follow-up, years	8 [6-11]	10 [8-12]	9 [6-11]
Functional renal grafts at 10 years post-transplantation, n (%)	296 (32)	162 (53)	458 (37)
Recipient gender, M/F (%)	567/363 (61/39)	198/109 (65/35)	765/472 (62/38)
Recipient age, years	56 [45-63]	50 [39-58]	54 [43-63]
Donor age, years	50 [38-58]	48 [38-58]	50 [38-58]
First transplantation, n (%)	788 (85)	230 (75)	1018 (82)
Pretransplant NDSA, n (%)	142 (15)	97 (32)	239 (19)
Serum creatinine at M12, µM	126 [105-155]	126 [103-152]	126 [105-154]
Proteinuria at M12, g/L	0.15 [0.11-0.23]	0.05 [0.05-0.13]	0.14 [0.10-0.21]
Proteinuria at M12, g/24h	MD	0.00 [0.00-0.35]	0.00 [0.00-0.35]
Return to dialysis, n (%)	136 (15)	86 (28)	222 (18)
Death with a functional graft (censured data), n (%)	MD	45 (15)	45 (15)
de novo NDSA, n (%)	MD	139 (45)	139 (45)
Time to onset, years	MD	4 [1-9]	4 [1-9]
de novo DSA, n (%)	46 (5)	67 (22)	113 (9)
Time to onset, years	4 [2-7]	5 [2-9]	5 [2-7]
Patients with onset of dnDSA in the first year after transplantation, n (%)	7 (14)	8 (17)	15 (16)
First acute rejection episode, n (%)	381 (41)	151 (49)	532 (43)
Time to onset, years	0.2 [0.0-1.0]	0.7 [0.3-1.6]	0.3 [0.0-1.0]

The three creatinine clusters allocated based on Scr at M1, M3, M6 and M12 are presented in Figure 1 and the corresponding survival curves in Figure 2. Graft loss-free survival was significantly associated with Scr clustering (log-rank test, $P < 0.001$). As expected, the patients of cluster A (53 %), with stable and low Scr, had the best graft survival probability. The patients of cluster B (43 %) with higher but stable creatinine had lower survival probabilities, mainly after 3 years post-transplantation. In cluster C (4 %), patients had very high Scr at M1 (cluster center at 429 µM) followed by a higher Scr plateau than the other clusters, and the worst graft survival. At 8 years post-transplantation, graft survival was 91%, 83% and 53% for clusters A, B and C, respectively.

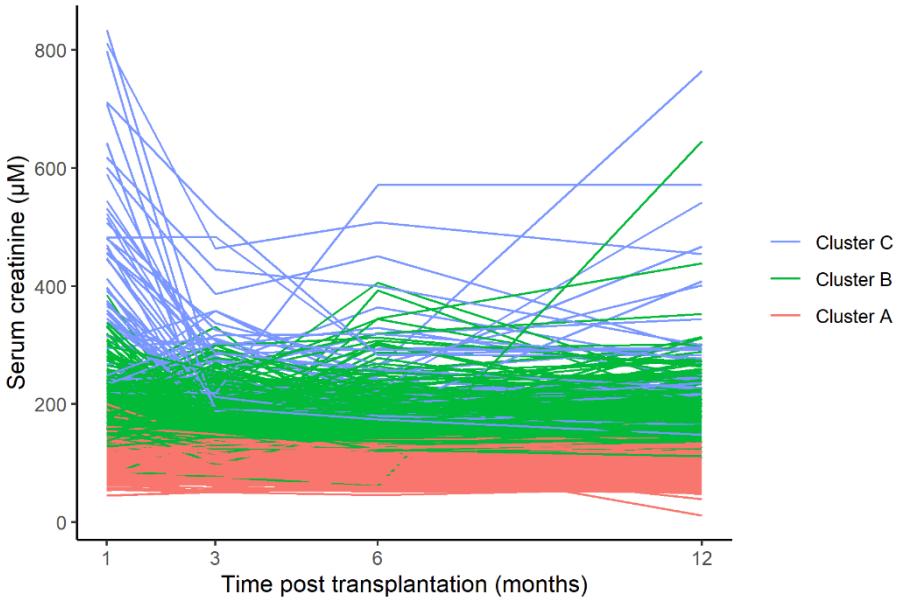


Figure 1. Serum creatinine profiles over months 1 to 12 post-transplantation, and their cluster attribution.

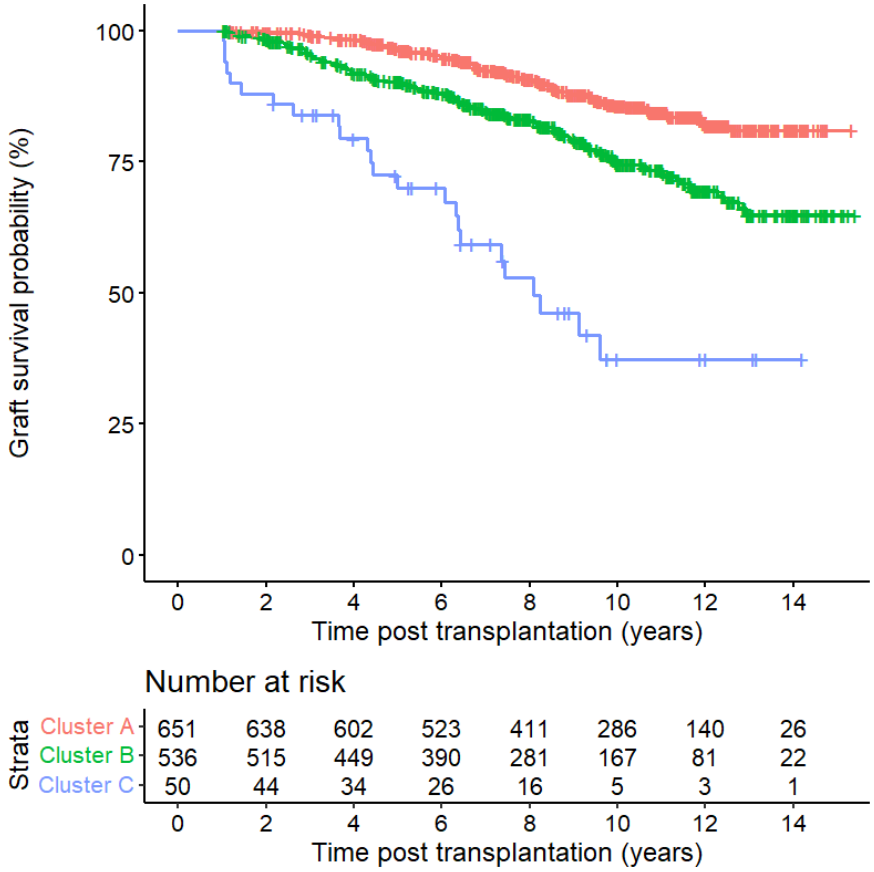


Figure 2. Comparison of Kaplan-Meier graft survival curves for the three serum creatinine clusters.

Graft survival after transplantation was significantly different across the three clusters ($p < 0.001$).

In a second step, AdGFS was calculated for each patient to classify them in four groups of low, intermediate, high, and very high risk (46 %, 34 %, 16 %, and 4 % of the patients, respectively). The AdGFS median values (interquartile range) was 2 (0-4). Figures 3 and 4 show the survival probabilities in these 4 groups for all patients, and for each center, respectively. For KU Leuven, HCL Lyon, and all the patients, there was a statistically significant difference in graft survival between the four risk groups (log-rank test, $P < 0.001$ for all three). Compared with the low-risk group, patients in the merged database had 1.9, 4.1 and 16.1 times more risk of losing their graft if they belonged to the intermediate, high, and very high-risk groups, respectively. At 8 years after transplantation, graft survival was 92 %, 89 %, 77 % and 36 % in the low, intermediate, high and very high-risk groups, respectively (Figure 3). At 10 years, graft survival was 89 %, 81 %, 59 % and 23 %, respectively. Detailed results of graft survival for each value of AdGFS are also presented as Supporting information (S2 Figure).

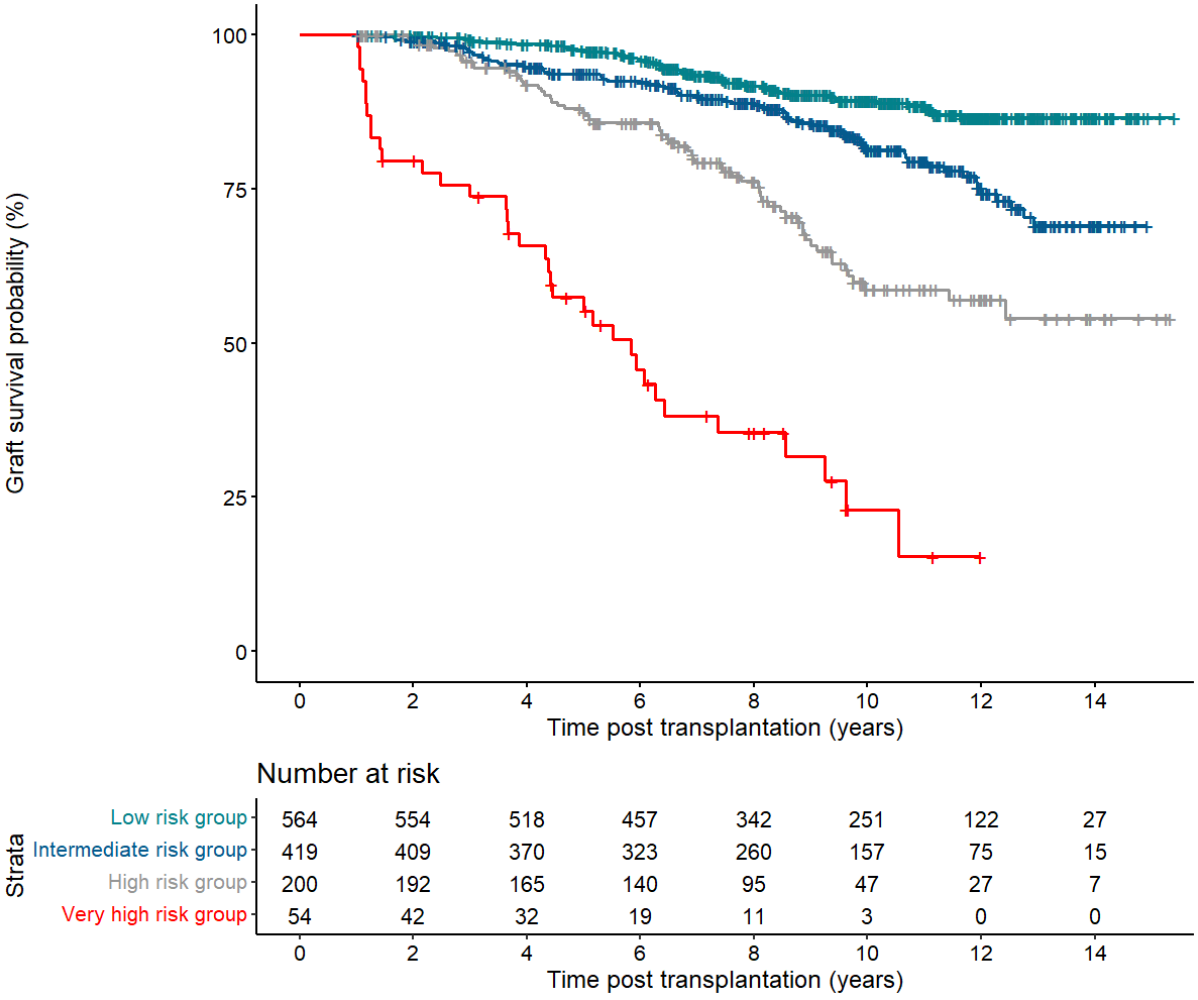


Figure 3. Kaplan-Meier graft survival curves in patients of the low, intermediate, high, and very high risk groups in the Leuven and Lyon datasets (pooled).

Patients were stratified according to the calculated score value: low risk (0), intermediate risk (2 or 4), high risk (6 or 8), and very high risk (10 or 12).

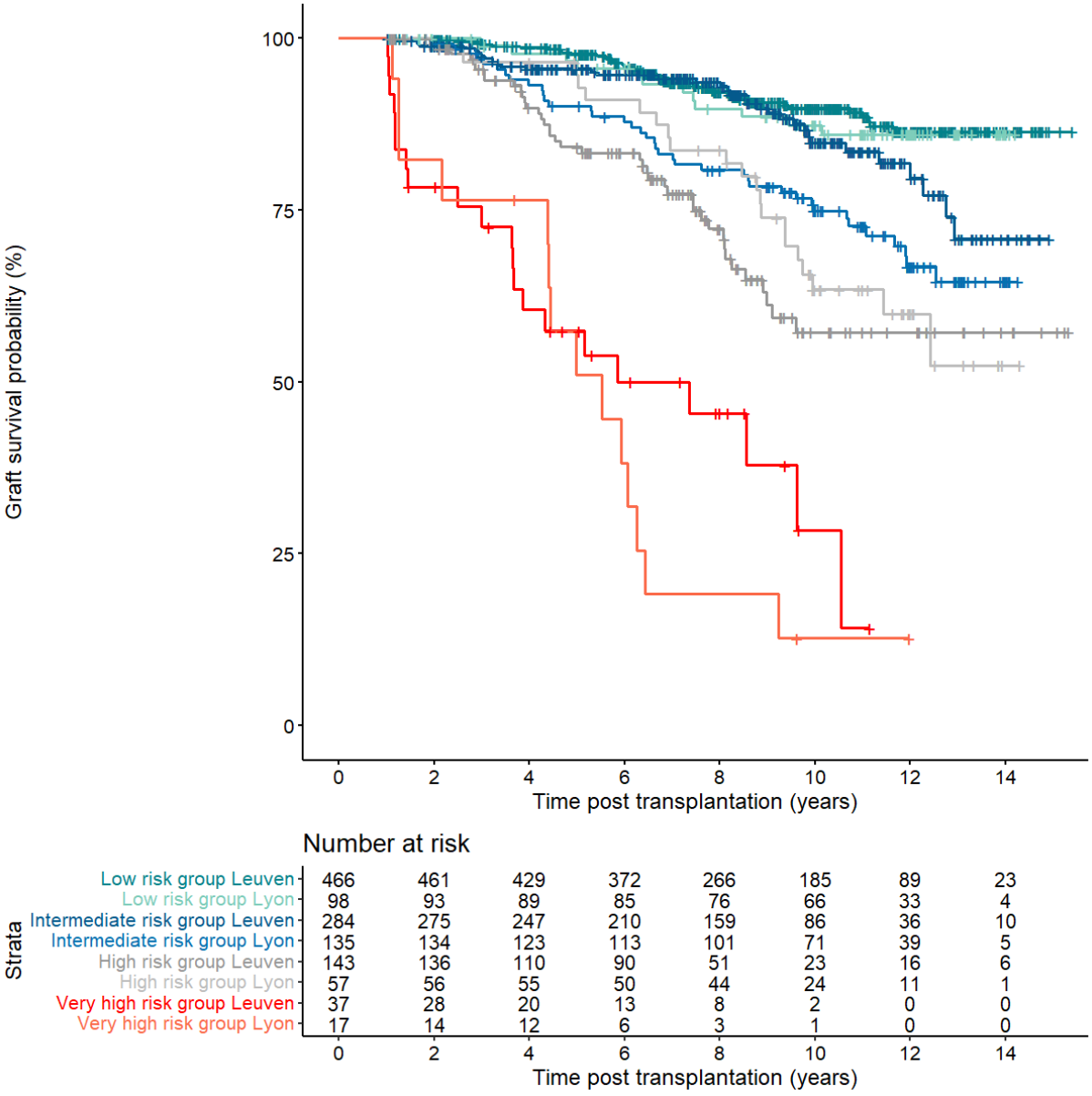


Figure 4. Kaplan-Meier graft survival curves in patients of the low, intermediate, high, and very high risk groups in the KU Leuven dataset (darker colors) and in the HCL Lyon dataset (lighter colors).

Patients were stratified according to the calculated score value: low risk (0), intermediate risk (2 or 4), high risk (6 or 8), and very high risk (10 or 12).

Figure 5 shows in more detail the predictive performance of the score depending on the time horizon. The ROC AUC (95% CI) was 0.86 (0.76-0.96), 0.76 (0.70-0.82) and 0.73 (0.69-0.77) at 2, 5 and 10 years post-transplantation, respectively.

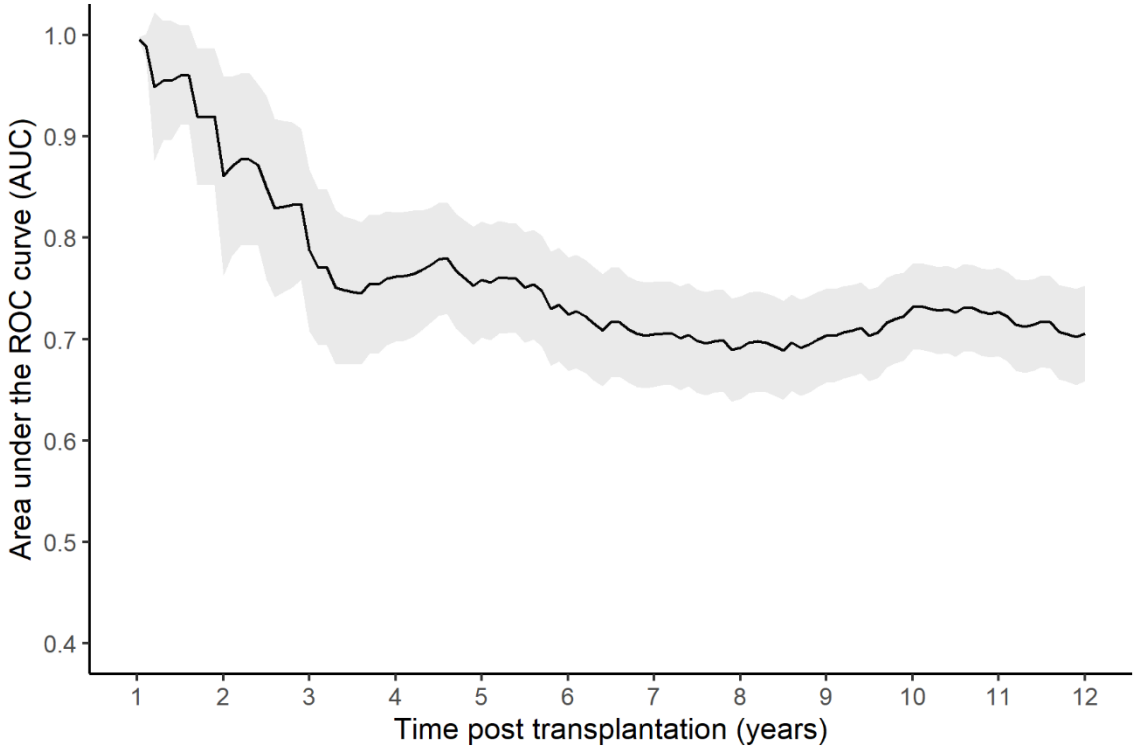


Figure 5. Areas under the ROC curve of AdGFS for prediction of graft failure (AdGFS) depending on the time post-transplantation.

The grey area is the 95 % confidence interval.

The performance characteristics of graft survival prediction are shown in Table 2 at different post-transplantation times, and for different AdGFS values. For example, a patient with a low score (AdGFS ≤ 2) has a probability of graft survival up to 10 years post-transplantation of approximately 90 % (negative predictive value or NPV). In contrast, a patient with a high score (AdGFS ≥ 10) has a probability of graft loss at or before 10 years post-transplantation of approximately 86 % (positive predictive value or PPV).

Table 2. Performance characteristics of the adjustable graft failure score (AdGFS) for cutoffs of 0, 2, 4, 6, 8, 10 and for different time horizons up to 10 years post-transplantation.

A low cutoff can be chosen to exclude graft loss with near certainty (to favor sensitivity and the negative predictive value), and a high cutoff to assert graft loss with near certainty (to favor specificity and the positive predictive value). Consequently, sensitivity with a high cutoff and specificity with a low one are knowingly unfavoured, and only provided (in gray) for completeness.

Score value	Censored post-transplantation time (years)	Sensitivity (%)	Negative Predictive Value (%)	Specificity (%)	Positive Predictive Value (%)	Accuracy (%)
> 0	2	94.0	99.8	46.3	2.4	46.8
	4	87.0	98.5	47.7	8.4	49.5
	6	79.2	95.9	48.7	13.3	51.0
	8	72.9	91.6	48.3	18.7	51.9
	10	76.6	89.9	54.8	30.8	57.7
> 2	2	94.0	99.8	54.6	2.8	55.2
	4	80.5	98.1	55.9	9.1	57.2
	6	73.3	95.5	56.4	14.4	58.2
	8	69.1	91.8	56.6	20.6	59.0
	10	72.1	89.6	63.1	34.0	64.0
> 4	2	70.4	99.5	80.5	4.8	80.1
	4	52.9	96.9	81.8	13.8	79.8
	6	50.5	94.4	83.1	22.9	79.7
	8	46.1	90.6	85.0	33.4	78.9
	10	45.6	86.2	89.1	52.4	79.3
> 6	2	64.5	99.5	93.0	11.4	92.5
	4	41.6	96.7	94.3	28.6	91.3
	6	36.0	93.7	95.3	43.4	89.7
	8	31.4	89.6	96.3	58.2	87.1
	10	24.0	83.0	97.6	72.4	82.2
> 8	2	64.5	99.5	96.5	20.5	96.1
	4	28.5	96.1	97.1	34.8	93.4
	6	25.5	92.9	98.0	55.7	91.3
	8	19.4	88.2	98.4	67.1	87.3
	10	15.1	81.6	99.3	85.9	81.8
> 10	2	23.4	98.9	99.4	36.0	98.3
	4	11.1	95.3	99.7	68.8	95.0
	6	7.4	91.5	100.0	100.0	91.6
	8	4.8	86.6	100.0	100.0	86.7
	10	3.2	79.7	100.0	100.0	79.9

Discussion

In this study, we validated in a large, independent and more recent database our previously published adjustable score for the prediction of kidney graft failure in the long term, named AdGFS [10]. AdGFS is an easy-to-use tool using 2-7 variables collected at one-year post-transplant and readily available in all transplantation centers (even those not performing systematic biopsies). These variables were compiled in a decision tree after being selected and combined using a Machine Learning method relying on decision trees, which has probably contributed to the performance of the score. AdGFS uses: donor age, pretransplant NDSA, the kinetic profile of Scr during the first year post-transplantation (inferred from Scr at M1, M3, M6, and M12), Scr and proteinuria at one-year post-transplantation, occurrence of DSA, occurrence of acute graft rejection. Analysis of graft survival at 8 years and 10 years post-transplantation showed very large and significant differences between the 4 risk groups.

This study demonstrates the robustness of AdGFS. For example, clusters A, B and C of Scr profiles represented 53/43/4 % of patients here, as compared to 57/37/6 % in the initial work [10]. At 10 years post-transplantation, graft survival probability was similar for clusters A and B (approximately 90/80%), but lower for cluster C in our study (< 50 % here vs. > 60 % in our previous report). As this cluster is always the smallest, differences in proportions may appear larger, but actually many events were recorded for this cluster just after the first transplant anniversary. The survival curves of the two studies were very similar. For both, the difference in survival probability between the “low” and “intermediate” risk groups was more pronounced from the 8th year post-transplantation onwards. The ROC AUC (95 % CI) was slightly lower at 10 years after transplantation in this study than in the external validation dataset of the initial work: 0.73 (0.69-0.77) and 0.79 (0.74-0.84), respectively.

The intermediate- and high-risk groups of patients (AdGFS = 2-4 and 6-8, respectively) followed-up in HCL Lyon had similar survival curves, crossing one another before 8 years post-transplantation. The small size of the high-risk group in Lyon (n = 57), more sensitive to incidental time variations, may be the reason for the slightly better survival probabilities before 8 years post-transplantation and more events thereafter, as a kind of catching up.

As can be seen in Table 2, a low AdGFS cutoff is not suitable to assert patient return to dialysis with near certainty, nor a high cutoff to exclude it (in grey). In the same way, it is logical that PPV (and specificity) is higher at 10 years post transplantation: the more time passes, the higher the probability that the graft will fail. It is reciprocal with NPV (and sensitivity) with early dates post-transplantation. Accordingly, users may not want to quantify the risk (based on the AdGFS value), but rather to be provided with a unique threshold to predict middle- to long-term return to dialysis. In this case, the best option is an intermediate cutoff of 4 or 6.

Not many prognostic models in general are externally validated [14], whereas it is a necessary step to demonstrate their reproducibility and generalizability to other types of patients, for instance recruited at other centers [11]. When actually done, the number of patients/events was sometimes small and/or the follow-up not long enough to conclude on their applicability for long-term patient care [15].

Prognostic models that aim to improve the prediction of clinical events are increasingly being developed and published in the field of kidney transplantation. Contrary to other scores, ours requires only 2-7 variables, not including biopsy data, and has been validated herein on a large independent cohort of patients followed up for a median (interquartile range) of 9 (6-

11) years. In contrast, the “dynamic, integrative system for predicting outcome” (DISPO) [16] requires at least 11 variables, including Banff criteria that are only available if biopsies are carried out. This system was developed and validated using several clinical databases of kidney graft recipients followed-up for an average of 7 (4-10) years, and the online application prognosticates survival probabilities up to 9 years, with 3 decimal digits and no confidence interval [17]. Unfortunately, the functionality of adding new follow-up evaluations (e.g. estimated glomerular filtration rate, eGFR) does not work [17]. Moreover, this predictor functions as a black box with no quantitative explanation of the contribution of each input data to the prediction, while AdGFS is a transparent, weighted decision tree. We compared the performance of the two predictors using information from *the first year* of follow-up after the biopsy (DISPO) or *the first year* post-transplantation (AdGFS), in the *European population*, to predict the event 5 years later and found ROC AUC = 0.80 and 0.72, respectively. When the event happened 6 years later, AUC was 0.76 with DISPO. Longer outcomes were not investigated. To reach this performance, DISPO requires 11 variables including information from biopsy reports, in addition to *all* the eGFR and proteinuria values over one year. This requires extracting all this information from the laboratory IT systems either automatically if permitted by the IT system, or manually, thus increasing the risk of human error. In comparison, AdGFS requires 2-7 variables (sometimes including 4 eGFR values) which can easily be typed-in by any physician, whatever the local IT system(s). All other comparisons between the two predictors would be biased, as the only performances shown in Raynaud et al.’s study concern predictions made using data accumulated over 2, 3, 4, 5 years, to predict an event closer in time: 4, 3, 2, 1 years later in the future, respectively. In the end, the main result presented by the authors of this paper is the average of these 5 ROC AUC values (as explained in the Supplementary appendix), inflating the result in a favorable way.

Loupy et al. [18] proposed another score called the iBox. Unfortunately, we could not test it is as the website link is corrupt [19]. An overview of its use in practice is available in the Supplementary appendix of the publication. Similarly to DISPO, the iBox score requires 10 variables including information from the first biopsy report. Many of the first biopsies occurred after the first-year post transplantation (up to 8 years, as presented in the Supplementary appendix). Therefore, making mid- or long-term predictions using mid- or long-term data is much easier, resulting in favorable metrics. For a new patient, the influence of each value is not shown. The AUC ROC of the iBox was not presented at different time horizons, contrary to the present study. Instead, the author chose the Harrell’s concordance index which is known to be biased upwards if the amount of censoring is high [20]. In the European cohort used for its validation, because of shorter follow-ups, the amount of censoring was indeed extremely high (90 - 97%).

In this study we validated AdGFS in a large, independent cohort and confirmed its clinical or research applicability, at least in Europe, for patient stratification regarding the risk of graft loss. It can be implemented on any portable device and provides, as early as the first transplantation anniversary, an adjustable *long*-term prediction of graft survival, for up to 9 years later.

Acknowledgments

The authors thank Ms. Karen Poole for manuscript editing.

Author Contributions

Conceptualization: PM AP ML.

Formal analysis: ML AP JBW.

Investigation: PM ML.

Methodology: ML AP JBW.

Writing - original draft: ML PM.

References

1. Evans RW, Manninen DL, Garrison LP, Hart LG, Blagg CR, Gutman RA, et al. The quality of life of patients with end-stage renal disease. N Engl J Med. 1985;312: 553–559. doi:10.1056/NEJM198502283120905

2. Wolfe RA, Ashby VB, Milford EL, Ojo AO, Ettenger RE, Agodoa LY, et al. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. N Engl J Med. 1999;341: 1725–1730. doi:10.1056/NEJM199912023412303

3. Alexander JW, Bennett LE, Breen TJ. Effect of donor age on outcome of kidney transplantation. A two-year analysis of transplants reported to the United Network for Organ Sharing Registry. Transplantation. 1994;57: 871–876.

4. Staeck A, Khadzhynov D, Kleinsteuber A, Lehner L, Duerr M, Budde K, et al. Influence of pretransplant class I and II non-donor-specific anti-HLA immunization on immunologic outcome and graft survival in kidney transplant recipients. Transpl Immunol. 2020;63: 101333. doi:10.1016/j.trim.2020.101333

5. Hariharan S, McBride MA, Cherikh WS, Tolleris CB, Bresnahan BA, Johnson CP. Post-transplant renal function in the first year predicts long-term kidney transplant survival. Kidney Int. 2002;62: 311–318. doi:10.1046/j.1523-1755.2002.00424.x

6. Naesens M, Lerut E, Emonds M-P, Herelixka A, Evenepoel P, Claes K, et al. Proteinuria as a Noninvasive Marker for Renal Allograft Histology and Failure: An Observational Cohort Study. J Am Soc Nephrol. 2016;27: 281–292. doi:10.1681/ASN.2015010062

7. Wan SS, Chadban SJ, Watson N, Wyburn K. Development and outcomes of de novo donor-specific antibodies in low, moderate, and high immunological risk kidney transplant recipients. Am J Transplant. 2020;20: 1351–1364. doi:10.1111/ajt.15754

8. Zhang R. Donor-Specific Antibodies in Kidney Transplant Recipients. Clin J Am Soc Nephrol. 2018;13: 182–192. doi:10.2215/CJN.00700117

9. Clayton PA, McDonald SP, Russ GR, Chadban SJ. Long-Term Outcomes after Acute Rejection in Kidney Transplant Recipients: An ANZDATA Analysis. J Am Soc Nephrol. 2019;30: 1697–1707. doi:10.1681/ASN.2018111101

10. Prémaud A, Filloux M, Gatault P, Thierry A, Büchler M, Munteanu E, et al. An adjustable predictive score of graft survival in kidney transplant patients and the levels of risk linked to de

novo donor-specific anti-HLA antibodies. PLoS One. 2017;12: e0180236. doi:10.1371/journal.pone.0180236

11. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14: 49–58. doi:10.1093/ckj/sfaa188

12. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005;61: 92–105. doi:10.1111/j.0006-341X.2005.030814.x

13. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol. 2017;17: 53. doi:10.1186/s12874-017-0332-6

14. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68: 25–34. doi:10.1016/j.jclinepi.2014.09.007

15. Ramspek CL, de Jong Y, Dekker FW, van Diepen M. Towards the best kidney failure prediction tool: a systematic review and selection aid. Nephrol Dial Transplant. 2020;35: 1527–1538. doi:10.1093/ndt/gfz018

16. Raynaud M, Aubert O, Divard G, Reese PP, Kamar N, Yoo D, et al. Dynamic prediction of renal survival among deeply phenotyped kidney transplant recipients using artificial intelligence: an observational, international, multicohort study. Lancet Digit Health. 2021;3: e795–e805. doi:10.1016/S2589-7500(21)00209-0

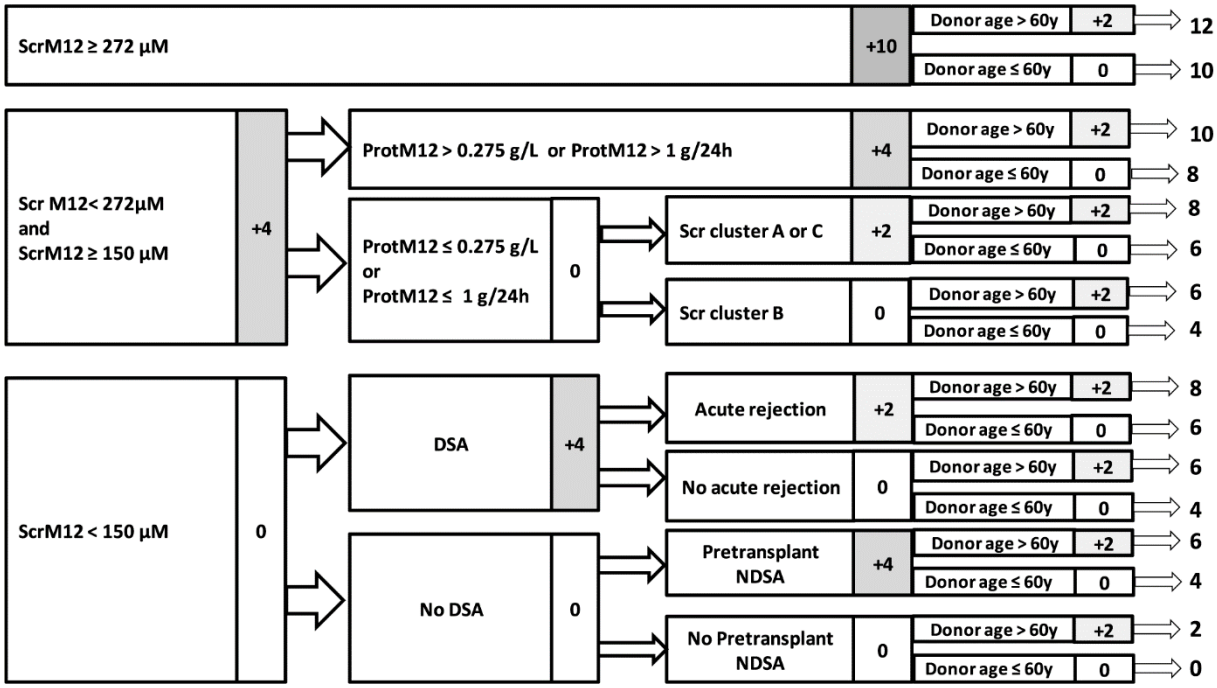
17. [cited 16 Sep 2022]. Available: https://transplant-prediction-system.shinyapps.io/AI-DISPO/

18. Loupy A, Aubert O, Orandi BJ, Naesens M, Bouatou Y, Raynaud M, et al. Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. BMJ. 2019;366: l4923. doi:10.1136/bmj.l4923

19. paristransplantgroup.org - Actualités du football mises à jour, derniers numéros de maillot des joueurs 2023. 16 May 2023 [cited 18 May 2023]. Available: https://www.paristransplantgroup.org/

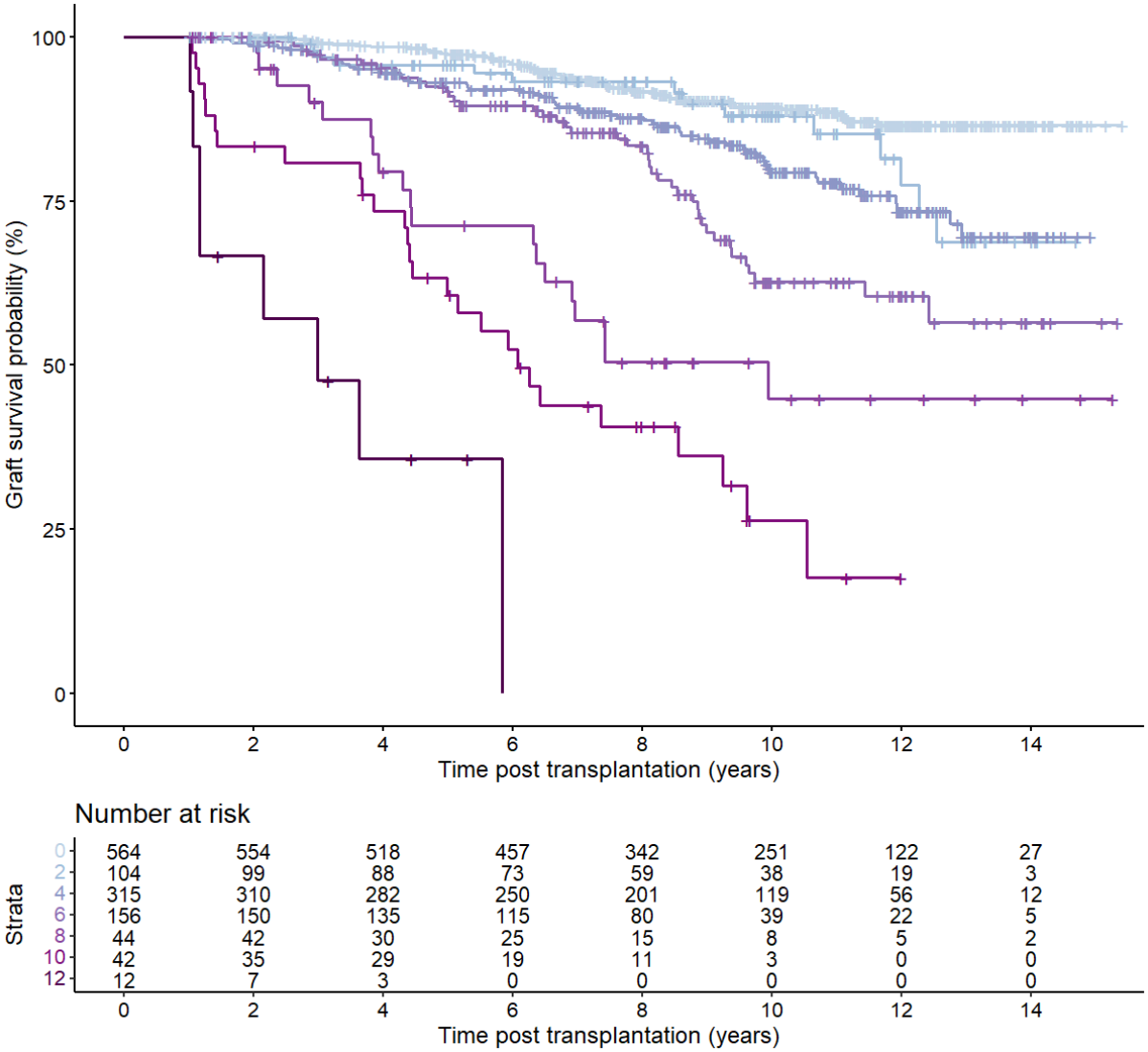
20. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30: 1105–1117. doi:10.1002/sim.4154

Supporting information



S1 Figure. Scoring system for computing AdGFS values (Prémaud et al., PLOS One, 2017).

ScrM12 = serum creatinine at 12 months post-transplantation. ProtM12 = proteinuria at 12 months post-transplantation. Scr = serum creatinine. dnDSA = de novo donor-specific anti-HLA antibodies. NDSA = non donor-specific anti-HLA antibodies.



S2 Figure. Comparison of Kaplan-Meier graft survival curves for each value of the AdGFS in the Leuven dataset and the Lyon dataset (pooled).

S1 Table. Clusters Centers.

For a new serum creatinine profile, the cluster can be allocated using these *clusters centers* and the function affectIndivC from the R package “kml”.

	1st month	3rd month	6th month	12th month
Cluster A	107.0	105.8	103.9	104.3
Cluster B	165.6	166.7	169.9	169.3
Cluster C	429.0	249.4	239.5	260.8

Discussion générale

Au cours de cette thèse, différents types d'outils de modélisation, d'estimation et de prédiction ont été employés dans le contexte de la pharmacologie de la transplantation rénale au sens large.

Dans le premier article, les AUC à l'origine des adaptations individuelles de posologie étaient obtenues par méthode bayésienne à partir d'un modèle PopPK. Les posologies conseillées étaient alors évaluées sur leur capacité à atteindre des cibles d'AUC. Cette étude peut donc être considérée comme l'évaluation prospective d'un outil d'intelligence artificielle déjà ancien (estimation bayésienne) sur la base d'un critère de jugement intermédiaire (atteinte des cibles d'AUC). L'idéal est de valider le suivi thérapeutique pharmacologique des immunosuppresseurs en utilisant d'autres critères de jugement comme la survenue de rejets aigus [146] ou la perte de fonction du greffon (retour en dialyse) [34], et ce, à grande échelle. C'est l'objectif du projet Rexetris[14] [29] présenté dans les Perspectives. Malheureusement, le devenir du patient greffé n'est pas une information disponible sur le site internet ABIS, bien que le centre demandeur ait la possibilité de préciser l'indication de chaque demande (rejet, suspicion de rejet, effets indésirables...). L'AUC est toutefois considérée comme un critère de jugement intermédiaire, dont les cibles pour réduire le risque de rejet sont consensuelles [5], ont été validées par un essai clinique randomisé chez les adultes [147] et ne seraient pas différentes chez les enfants [148,149].

Pour définir les cibles d'AUC des *autres* immunosuppresseurs, deux conditions sont nécessaires. Il faut être capable de concevoir des modèles robustes capables d'estimer les AUC à partir d'un nombre limité de prélèvements. Et la survie du greffon doit être étudiée à court et long terme, pour fixer des objectifs d'exposition avec un niveau de preuve suffisant. En outre, pour surveiller les AUC en pratique courante, les procédures de prélèvement doivent être acceptables pour les patients et leur médecin transplanteur. Un certain nombre de centres de transplantation rénale français postent sur le site ABIS des demandes conjointes d'estimation de l'AUC du MMF et du CNI associé (le plus souvent tacrolimus). Ce suivi est programmé et systématique pour de nombreux patients (voire tous). Le ML peut rendre encore plus accessible le suivi thérapeutique pharmacologique par l'AUC. Avec les mêmes performances que les estimateurs du MAP, l'algorithme XGboost était capable d'estimer l'AUC de tacrolimus à partir de 2 prélèvements seulement [133].

Dans un deuxième travail, nous avons mis au point un outil de ML pour l'estimation de l'AUC d'évérolimus dans l'espoir de pouvoir réduire le nombre de prélèvements sanguins nécessaires. En effet, dans une étude à laquelle j'ai participée, Woillard et al. avions entrainé un modèle d'estimation d'AUC du tacrolimus sur des profils réels. Avec 3 concentrations, les erreurs RMSE normalisées variaient de 7,3 à 11,5 % [133]. Avec 2 concentrations seulement, les performances étaient toujours très satisfaisantes : RMSE normalisées de 9,0 à 12,9 %. En revanche, l'entraînement d'un algorithme XGBoost sur des profils d'*acide mycophénolique* issus de patients réels [116] n'avait pas permis de réduire le nombre de prélèvements malgré une base de données importante (n = 12 877 profils réels). En utilisant 3 concentrations, les

[14] Rexetris : Relations EXposition - Effet à long terme chez le Transplanté Rénal des médicaments ImmunoSuppresseurs.

erreurs RMSE normalisées variaient de 14,7 à 22,5 % sur 4 jeux de données indépendants, et avec 2 concentrations elles variaient entre 17,8 et 26,3 %. Pour l'*évérolimus*, les performances n'étaient pas satisfaisantes non plus quand seulement 2 valeurs de concentrations étaient incluses.

Quand nous avons utilisé pour la première fois du ML pour estimer les AUC de tacrolimus à partir de 3 concentrations, nous avions un nombre très important de profils réels : n = 1452-4997[15] [133]. Les performances étaient meilleures avec XGBoost (RMSE normalisées entre 7,3 et 11,5 %) qu'avec un estimateur du MAP (RMSE normalisées entre 9,1 et 15,7 %). Puis, 4192 simulations ont été utilisées seules pour créer un nouveau modèle d'estimation d'AUC pour le tacrolimus [134]. Les RMSE normalisées variaient alors de 8,0 à 9,1 % en fonction du type de transplantation sur les bases de données externes, avec une différence importante par rapport au jeu de données de test (RMSE normalisé à 2,7 %). Avec l'évérolimus, nous avons commencé par utiliser autant de profils simulés que de profils réels. Puis nous avons progressivement augmenté le nombre de profils simulés dans le jeu de données d'entraînement pour améliorer les performances du modèle, en s'arrêtant à 15,000. Au final, *l'apport des simulations* pour améliorer l'apprentissage du modèle de ML n'était pas significatif pour le tacrolimus (RMSE normalisé proche de 9 % dans les deux cas) alors que cette différence était importante dans notre publication : 17,2 % versus 10,3 %, avec les 508 profils réels et les 5016 simulations, respectivement.

Avec une approche différente, Ponthier et al. ont publié d'autres modèles d'estimation d'AUC à partir, entre autres, de 3 concentrations [150]. Leur publication concernait le MeltDose®-tacrolimus.[16] Les profils de concentrations utilisés *pendant l'apprentissage* étaient cette fois-ci *complets*, permettant ainsi de fournir des AUC de référence de plus grande qualité pour cette étape. La contrepartie était que le nombre d'AUC de référence était limité à 210. L'imprécision RMSE normalisée était de 10,1 et 9,1 % sur des jeux de données indépendants de transplantation rénale (n = 16) et hépatique (n = 48), respectivement. Cet article avait permis de montrer que les performances (en validation externe) étaient similaires et pas meilleures que celles obtenues en utilisant un jeu de données d'apprentissage plus large d'estimations moins précises de l'AUC. En effet, dans la précédente étude, 1452-4997 profils issus du site internet ABIS avaient été utilisés [133].

Dans un autre article, Ponthier et al. ont développé un modèle d'estimation a priori de l'AUC de vancomycine [151] pour les nouveau-nés prématurés, en se basant sur des simulations issues d'un modèle de PopPK, et surtout en comparant plusieurs algorithmes de ML (XGBoost, GLMNET,[17] MARS[18]). Les RMSE normalisées variaient de 38,1 à 39,8 % en fonction de l'algorithme employé, le meilleur étant XGBoost. Les variables explicatives utilisées pour la prédiction étaient : la dose de charge, la dose administrée en perfusion continue, l'âge gestationnel, le poids à la naissance, le délai entre la naissance et la survenue

[15] Les profils étaient issus de patients réels traités par tacrolimus avec 1 prise par jour (n = 1452) ou 2 prises par jour (n = 4997).

[16] MeltDose®-tacrolimus : Envarsus®.

[17] GLMNET : package du logiciel R utilisé pour ajuster des modèles de régression linéaire et de régression logistique avec les régularisations L1 et L2.

[18] MARS : régression multivariée par spline adaptative.

de l'infection, l'âge post-menstruel[19] au moment de la survenue de l'infection, le poids gagné par jour, le poids actuel, et la créatininémie.

En résumé, entraîner des algorithmes de prédiction de l'AUC de médicaments à partir de profils pharmacocinétiques riches ne semble pas plus efficace que de les entraîner sur des AUC estimées par méthode bayésienne MAP, à condition probablement que ces dernières ne soient pas biaisées (ce qui est le cas, nous l'espérons, de celles développées dans notre équipe depuis 25 ans). Entraîner les algorithmes sur des données simulées semble présenter plusieurs avantages : *enrichir* une base d'entraînement de patients réels avec un modèle PopPK indépendant, obtenir un nombre *élevé* de profils pharmacocinétiques pour l'apprentissage d'un algorithme de ML, et générer des profils avec des caractéristiques *extrêmes* (dose, poids, etc.). Les simulations pourraient aussi permettre de quantifier l'influence d'une petite variation dans les temps de prélèvement. Quand un modèle de la littérature existe déjà, les simulations pourraient également éviter de répéter des études pharmacocinétiques (coût humain et financier) : prises de sang répétées, mesure de nombreuses concentrations. A l'inverse, les inconvénients des simulations sont : le risque de biais si le modèle PK initial est mal formulé ou que ses paramètres ont été ajustés sur une population non représentative de la population d'intérêt ; le risque de faire du surapprentissage quand les algorithmes de ML sont validés sur des simulations issues du *même modèle* PopPK ; et le temps de calcul qui peut devenir rédhibitoire pour l'entraînement de l'algorithme de ML sur des trop grands jeux de données simulées. L'utilisation de simulations exige donc une validation de l'algorithme obtenu sur un ou plusieurs jeux indépendants de données réelles et riches (profils complets de patients réels). Comme montré dans ce travail, elle nécessite également de vérifier l'absence de surapprentissage en faisant varier le nombre de données simulées, et dans l'idéal d'identifier le nombre optimal.

Dans la littérature, nous n'avons malheureusement pas trouvé d'autres publications sur l'utilisation d'algorithmes de ML pour l'estimation de l'aire sous la courbe des concentrations de médicaments en général.

En pharmacologie clinique et thérapeutique, l'effet des traitements est au moins aussi important que leur pharmacocinétique. Le rejet aigu, caractérisé par une inflammation des tissus, est l'une des causes de perte de fonction du greffon rénal à long terme. C'est également un critère de jugement de substitution (*surrogate*) de l'efficacité des traitements immunosuppresseurs en transplantation d'organe. Le terme de *rejet aigu prouvé par biopsie* (ou BPAR) est une définition *non spécifique* qui empêche de distinguer les effets à long terme de certaines thérapies immunosuppressives, des effets à court terme sur des rejets non spécifiques [146,152]. La distinction entre les différentes formes de rejet est primordiale pour la définition du critère de jugement dans les essais cliniques, notamment pour évaluer l'efficacité de l'immunosuppression[20] (avis du Comité des médicaments à usage humain de l'Agence européenne des médicaments ou EMA).

Dans une étude récente, Kers et al. ont utilisé une méthode de Deep Learning (réseau de neurones convolutifs, ou CNN) pour lire des lames de biopsie rénale, et proposer directement un diagnostic *simplifié*. L'algorithme ne permettait donc pas de faire la distinction entre les différentes formes de rejet. Les 3 conclusions possibles pour chaque biopsie étaient :

[19] $\hat{A}ge\ post\ menstruel = \hat{a}ge\ gestationnel + l'\hat{a}ge\ chronologique$

[20] Par exemple, un rejet aigu non-spécifique (TCMR précoce) ne serait pas pertinent comme critère de jugement pour l'efficacité à long terme d'un médicament immunosuppresseur [153].

normale, rejet, ou autre maladie (néphropathies à polyomavirus[21]) [154]. Dans cette étude rétrospective, les CNN étaient ainsi entraînés à faire de la reconnaissance d'image à partir de lames numérisées. Les rejets étaient définis par un unique néphropathologiste expérimenté utilisant les règles de Banff. Les cas suspects, eux, étaient considérés comme « rejet ». Quand les modèles étaient validés sur des jeux de données indépendants, les performances étaient bonnes pour détecter *une* maladie (AUC ROC = 0,83). Mais quand il s'agissait de *distinguer* la catégorie « rejet » de la catégorie « autre maladie », la classification était proche d'un choix au hasard (AUC ROC = 0,50 à 0,76). Les auteurs ont donc reconnu qu'actuellement ces modèles n'étaient pas assez performants pour remplacer un anatomopathologiste humain. De plus, les modèles qui y étaient présentés avaient un effet black box important puisqu'il n'y avait pas d'étapes intermédiaires, où l'étendue de chaque type de lésion aurait pu être décrite. Néanmoins les images les plus prédictives ont pu être extraites automatiquement pour une analyse visuelle par l'opérateur. Cette publication illustre la difficulté actuelle d'une automatisation globale du diagnostic de rejet du greffon rénal.

Dans les 10 dernières années, la classification de Banff n'a pas changé de manière importante pour les diagnostics de rejet. Si la classification reste stable dans les prochaines années, il sera inutile de recommencer l'entraînement de notre algorithme, qui nécessiterait de redéfinir tous les diagnostics dans une base de données d'apprentissage suffisamment grande. Les modèles présentés ici n'avaient pas comme objectif d'utiliser la classification de Banff comme référence d'apprentissage. Au contraire, ils ont été entraînés sur l'expérience combinée de spécialistes, qui lisent des biopsies de greffons rénaux ou suivent quotidiennement des patients transplantés et participent à la publication de nombreux travaux sur la définition et la détection des rejets.

Si un nouvel apprentissage *différent* devait être envisagé, il pourrait utiliser les catégories de Banff comme diagnostic de référence pour chaque cas. Contrairement à la classification de Banff, l'entraînement (supervisé sur ces diagnostics) inclurait des données cliniques comme la créatininémie, la protéinurie des 24h (ou le rapport protéinurie sur créatininurie) ... A la différence de la classification de Banff, l'algorithme permettrait de classer les biopsies par un score continu de 0 à 1 qui tiendrait compte des informations cliniques pertinentes.

Pour prédire la survie du greffon à long terme, nous pourrions aussi utiliser les *scores* des lésions élémentaires des biopsies de la première année post transplantation. De cette manière, nous ne passerions pas par l'étape intermédiaire de l'*interprétation* des scores pour le diagnostic de rejet aigu, avant d'utiliser ce dernier comme prédicteur de survie du greffon. Néanmoins, l'utilisation du diagnostic, comme prédicteur supplémentaire, pourrait aider l'algorithme. En ML, il est en effet fréquent d'ajouter certaines variables transformées, en plus des prédicteurs disponibles de base, dans le but d'améliorer les prédictions (*feature engineering*).

Dans le futur, il serait intéressant d'évaluer prospectivement l'apport de l'utilisation de tels outils d'aide au diagnostic de rejet aigu, ou plus généralement, de lésions du greffon, pour le suivi du patient, grâce à un ou des essais comparatifs randomisés. Toutefois, cette validation clinique ultime n'a jamais été réalisée pour la classification de Banff initiale et ses variantes successives. L'utilisation d'essais comparatifs randomisés est un des nombreux exemples du

[21] Néphropathie à polyomavirus : inclut les néphropathies à BK virus.

progrès médical en termes de validation des outils diagnostiques (vision optimiste). Au contraire, ces exigences peuvent paraître injustifiées vis-à-vis des outils d'IA appliqués au diagnostic, lorsque leurs performances surpassent celles des meilleurs outils existants (vision pessimiste).

Dans le domaine de la transplantation rénale, les critères de jugements *principaux* reconnus par l'EMA sont : les rejets aigus, le DFGe, la perte de fonction du greffon, et le décès du patient. Les biomarqueurs moléculaires sont également étudiés comme critère de jugement secondaire. Ils nécessitent bien sûr d'être validés sur un nombre élevé de patients transplantés [155]. Par exemple, dans une biopsie, l'expression des gènes et l'histologie sont complémentaires. L'expression des gènes montre de manière précoce l'activité immunitaire en cours [156], tandis que l'histologie est un meilleur outil d'évaluation des lésions cumulées et des structures affectées. Dans certains centres européens, réaliser des biopsies systématiques pendant la 1re année transplantation est donc devenue la norme [157]. Ce geste reste néanmoins invasif et n'est pas dénué de possibles complications [158]. Un biomarqueur idéal est donc un biomarqueur totalement non-invasif, comme ceux présents dans l'urine [159,160]. Dans ce cas, comme pour les simples prélèvements sanguins, le risque de faux positifs par activation non spécifique des lymphocytes T doit être pris en compte : néphropathie à BK virus, infection des voies urinaires, infection à CMV [161]. A ce jour, il n'existe donc pas encore de biomarqueur aussi consensuel que le DFGe (calculé grâce à la créatininémie).

L'outil de ML que nous avons développé pour améliorer le diagnostic de rejet aigu du greffon rénal nous semble pouvoir être utile au clinicien et au patient à *court terme*. La classification fournit une aide à la décision pour déclencher ou non un traitement immunosuppresseur *curatif*[22] du rejet aigu : par exemple par des échanges plasmatiques, des fortes doses de corticoïdes ou de globuline antilymphocytaire [162]. L'arbre de décision, que nous avons validé dans une base de données indépendante pour prédire le pronostic à *long terme*, a pour objectif d'aider les cliniciens à adapter le suivi et le traitement du patient de manière *préventive*. En fonction du pronostic prédit, les adaptations peuvent concerner : la fréquence de la surveillance de la fonction du greffon, un allègement ou une intensification du suivi thérapeutique pharmacologique, un allègement ou une augmentation de l'immunosuppression de maintenance, etc... L'objectif est d'améliorer la balance bénéfice-risque du traitement immunosuppresseur d'entretien et d'améliorer la qualité de vie du patient.

[22] En réalité, ces traitements ne peuvent pas être complétement curatifs, ils vont surtout permettre d'interrompre la réaction immunologique aiguë et donc de ralentir la vitesse de dégradation de la fonction du greffon.

Perspectives

Les travaux présentés ici (et d'autres auxquelles j'ai participé [116,133,134,150]) ont porté de manière indépendante sur l'exposition aux immunosuppresseurs (Article 1 et Article 2), et sur les critères de jugement d'efficacité de ces médicaments les plus importants dans le contexte de la transplantation rénale, que sont le rejet et la perte de fonction du greffon (Article 3 et Article 4).

L'idéal serait de mettre en parallèle ces deux types informations (exposition et effets), pour de nombreux patients, et sur des suivis longs. C'est l'objectif du projet Rexetris, porté par le Pr. Pierre Marquet dans notre équipe, qui permettra d'étudier de manière rétrospective les Relations EXposition - Effet des médicaments ImmunoSuppresseurs à long terme chez le Transplanté Rénal [29]. Ce projet réunit au CHU de Limoges et à l'UMR1248 Inserm « Pharmacologie & Transplantation » des médecins, des pharmaciens, des data managers, et des ingénieurs en biostatistiques et IA. L'objectif est de rapprocher des données sur le suivi thérapeutique pharmacologique de patients transplantés, avec les évènements importants de leur suivi (effets indésirables, rejets, date de retour en dialyse, décès ...). Pour les transplantés rénaux qui ont pu en bénéficier (18 440 transplantations chez 17 426 patients différents entre le 01/01/2005 et le 31/12/2021), le site internet ABIS fournira des informations sur *l'exposition* aux immunosuppresseurs : concentrations résiduelles, AUC, et concentrations maximales estimées. Le devenir des patients pourra être étudié grâce à la base de données Cristal de l'Agence de la Biomédecine et à celle du Système National des Données de Santé (SNDS).

Cristal est un registre de tous les patients ayant bénéficié d'une greffe d'organe en France, alimenté par tous les professionnels de santé impliqués dans le prélèvement et la greffe d'organe, mais aussi les laboratoires HLA et les anatomopathologistes. Il collecte les informations nécessaires au suivi des patients transplantés, tout en garantissant l'anonymat entre le donneur et le receveur. Il contient des informations précieuses sur le donneur, la greffe, et le suivi médical du receveur après greffe, dont les complications et les rejets (n = 49 886 greffons transplantés à 47 842 receveurs entre le 01/01/2005 et le 31/12/2021). Les bases de données du SNDS, elles, regroupent des informations concernant 42 705 transplantés rénaux qui ont pu être appariés avec la base Cristal, sur :

- l'Assurance Maladie (base SNIIRAM[23]) : remboursements effectués par l'ensemble des régimes ;
- les hôpitaux (base PMSI[24]) : résumés des séjours ;
- les causes médicales de décès (base du CépiDC de l'Inserm).

Le Health Data Hub (HDH) [163] est la plateforme qui permet de fournir un accès à ces données de santé, après autorisation de la Commission nationale de l'informatique et des libertés (CNIL). La réception, le stockage, et le traitement des données par des logiciels statistiques se fait dans un « espace projet », grâce à des serveurs dédiés accessibles à distance.

[23] Système national d'information inter-régimes de l'Assurance maladie.

[24] Programme de médicalisation des systèmes d'information.

Grâce au chaînage de ces bases de données, nous pourrons ainsi dans un premier temps établir des modèles de risques *cliniques*. Nous comparerons :

- les caractéristiques « de base » du suivi du greffon : qualité de l'organe transplanté, compatibilité, déroulement de la greffe, informations démographiques générales et antécédents du receveur ;

- les données de survie des greffons, des patients, mais aussi les informations sur la fonction rénale, ou encore l'apparition de comorbidités sévères (cancer, MACE,[25] diabète).

Les outils de ML pourront nous aider à sélectionner les informations les plus prédictives sur la survie à long terme du greffon, parmi les dizaines de variables disponibles. Les déterminants *pharmacologiques* seront ensuite introduits dans les modèles de risques élaborés lors de l'étape précédente. Les stratégies immunosuppressives les plus représentées pourront être alors être comparées à profil de risque égal (Figure 25). De la même manière, plusieurs niveaux d'exposition pourront être comparés.

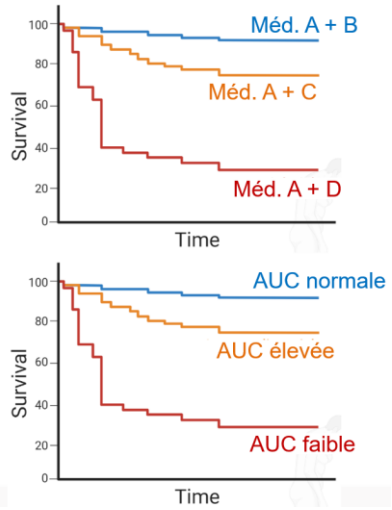


Figure 25 : Exemples de comparaisons possibles dans le projet Rexetris

Les différences de survie entre plusieurs stratégies thérapeutiques (en haut) ou plusieurs niveaux d'exposition (en bas) pourront être étudiées grâce : au calcul du hazard ratio par modèle de Cox, ou encore la visualisation par des courbes de Kaplan-Meier représentées ici.

Dans un deuxième temps, ces bases de données nous permettront d'évaluer l'efficacité à long terme de la stratégie d'adaptation de posologie développée par le CHU de Limoges depuis 2005. Un groupe de patients suivi sur ABIS (n = 17 426) sera ainsi au groupe témoin des patients n'en ayant pas bénéficié (n = 32 460). La différence de survie sans perte du greffon pourra alors être directement étudiée, à la différence de l'Article 1 dans ce travail, qui se concentrait sur un critère intermédiaire, l'atteinte des cibles d'AUC recommandées.

Concernant le projet Rexetris, l'accès tardif aux données pendant mon doctorat (janvier 2023 pour les données CRISTAL et fin juillet 2023 pour les données du SNDS) ne m'a pas permis de présenter des premiers résultats. Le lourd travail de nettoyage des données a

[25] Major adverse cardiovascular event.

néanmoins été bien avancé, et nous espérons débuter les analyses statistiques très prochainement. Je serai très impliqué dans les 3 types d'études mentionnées plus haut.

Dans les différents travaux présentés dans ce document, la médecine de précision est principalement utilisée dans un mode *adaptatif*. L'exposition à un médicament est mesurée (Article 2), ce qui va entraîner une proposition d'adaptation de posologie (Article 1). De la même manière, la prise en charge va être adaptée à court et long terme : à la suite de la constatation d'un diagnostic de rejet aigu (Article 3), ou alors après un bilan initial de fin de première année post transplantation (Article 4). Malheureusement, nous avons pu remarquer que le suivi thérapeutique pharmacologique, est souvent initié tardivement : par exemple 41 % des premières estimations d'AUC d'AMP chez les enfants étaient réalisées à plus de 1 an post-transplantation. Or, seulement 50% des premières estimations d'AUC étaient dans les cibles. Pire, dans de nombreux cas, après une première proposition d'ajustement de dose le contrôle de l'AUC est effectué tardivement. En parallèle, le rejet « aigu » est probablement un processus relativement lent et continu, qui commence à un niveau moléculaire pour n'être détecté que trop tard au niveau clinico-biologique, ou dans le meilleur des cas à un niveau histologique permettant de suspecter le début de lésions spécifiques.

Le programme de recherche Digphat[26] porté par le Pr. Jean-Baptiste Woillard dans notre équipe et le Dr. Christophe Battail de l'Université de Grenoble Alpes, et financé par le PEPR[27] Santé Numérique propose une vision différente de la médecine de précision qui ne serait plus uniquement utilisée *a posteriori* [164]. L'objectif est de rassembler des modèles développés à plusieurs échelles : moléculaire, cellulaire, et de l'organisme entier. Ces modèles mécanistiques mono-échelles permettront de produire des métamodèles grâce au ML. Ces derniers pourront ainsi améliorer les performances prédictives en termes de personnalisation des traitements, et ce, dans le domaine de la transplantation, l'oncologie, et l'antibiothérapie. Le but est de pouvoir évaluer un traitement particulier *a priori*, grâce à un jumeau numérique pharmacologique. Les métamodèles se baseront alors sur les caractéristiques individuelles d'un patient au niveau moléculaire (biomarqueurs), pharmacogénomique, démographique, afin de déterminer la probabilité qu'un traitement soit efficace. Chez un patient avec un pronostic moins favorable, c'est la probabilité que le traitement de deuxième ligne ou qu'une combinaison de traitements soit efficace qui pourra être estimée. De manière évolutive, le recueil de nouvelles données longitudinales permettra de réévaluer la probabilité de succès (ou d'effets indésirables) d'une stratégie thérapeutique. Ce projet nécessite des compétences transversales et complémentaires dans plusieurs domaines : pharmacologie de système, pharmacométrie, biostatistique et bioinformatique. Ce programme innovant profite donc de l'expertise et de la collaboration de plusieurs centres de recherche qui travaillaient jusqu'ici individuellement, et sur des modèles à une seule échelle. Je serai également associé à ces travaux, concernant la médecine personnalisée en transplantation.

[26] Digphat : Pharmacological Digital Twin.

[27] PEPR : Programme et Equipements Prioritaires de Recherche.

A travers les travaux réalisés pendant cette thèse, les étapes importantes de la création d'un modèle de ML ont été abordées :

- séparation du jeu de données pour sa validation
- imputation de données manquantes
- genèse éventuelle de données synthétiques
- transformation de certaines variables
- choix de l'algorithme de ML
- choix du seuil de positivité
- évaluation poussée des performances
- et explication de la prédiction à l'utilisateur.

Pour simplifier la prise de décision pendant tout ce processus, le laboratoire van der Schaar à Cambridge, UK, avec lequel notre équipe collabore, a développé un package nommé AutoPrognosis [165]. C'est un algorithme qui propose d'optimiser automatiquement la phase d'apprentissage en proposant un *pipeline* pour la création de modèles pronostiques. A chaque étape (imputation, transformation, classification...), plusieurs algorithmes sont disponibles, et le meilleur est choisi par méthode bayésienne (Figure 26). Ce package est actuellement disponible en ligne [166]. Bien qu'il s'adresse à des utilisateurs confirmés du logiciel Python, il permet aux professionnels de santé de simplifier le maniement de modèles de classification (ou de survie), comme d'autres packages connus sur R [166].

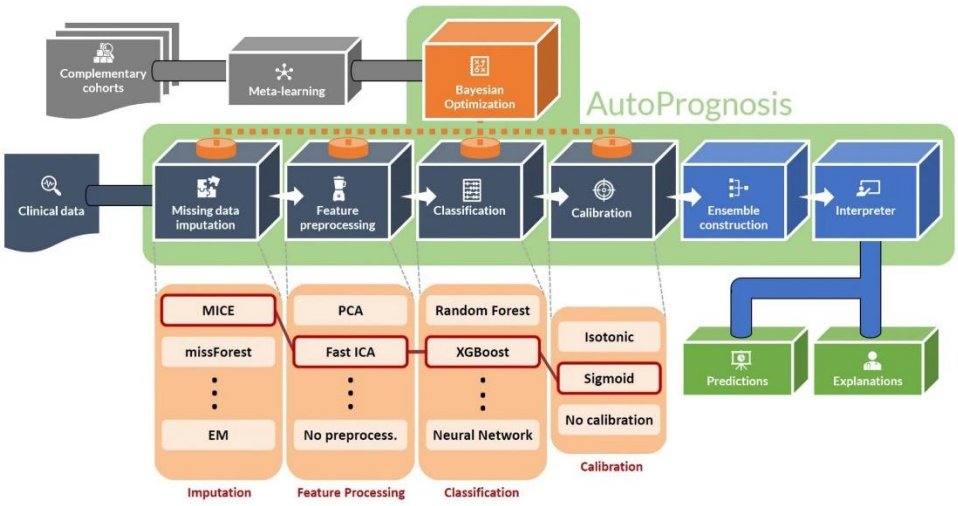


Figure 26 : Schéma de fonctionnement de l'outil AutoPrognosis

EM, expectation-maximization ; ICA, independent component analysis ; MICE, multiple imputation by chained equations ; PCA, analyse en composante principale.

Source : site internet dédié [145]

AutoPrognosis a déjà été testé avec succès dans plusieurs applications en santé : prédiction de maladies cardiovasculaires [167], survie des patients atteints de la

mucoviscidose [168], prédiction des résultats d'IRM préopératoires dans le cancer du sein (vrais positifs, faux positifs, faux négatifs, facteurs responsables) [169]. Un de mes travaux en cours est de le tester pour générer des modèles de survie du greffon.

A l'avenir, nous souhaiterions également développer notre propre package sur R pour l'application des règles de Banff. Cela permettrait d'utiliser cette classification sur des jeux de données de taille importante, en fonction des données disponibles, et des différentes versions qui ont existés. Grâce à ce type de package, les modèles de ML qui utilisent les diagnostics de Banff comme référence pourraient être actualisés régulièrement en fonction des mises à jour des catégories de Banff.

Quand un modèle de ML est créé et validé dans un jeu de validation externe, on peut envisager de l'utiliser en pratique clinique. Il existe alors 3 moyens différents pour son implémentation en routine.

La première option nécessite que le modèle soit facilement utilisable, sans nécessité d'utiliser un outil spécifique. Comme dans l'Article 4 ici, cela peut être un arbre de décision (ou un arbre de régression) entraîné informatiquement. Cela peut aussi être une équation, qui ne nécessiterait qu'une calculatrice pour être utilisée. Dans tous les cas, ces modèles peuvent être utilisés « sur un coin de bureau », sans laisser de traces nominatives. En général, une telle simplification d'un modèle de prédiction a un coût en termes de performances (Tableau 3). Mais il peut être facilement diffusable (affiche, publication d'article papier ou en ligne, livre), et le résultat facilement compréhensible, puisqu'il suffit de suivre toutes les branches de l'arbre, ou de lire les coefficients de la formule, pour comprendre son fonctionnement (rétro-ingénierie).

La deuxième option est de proposer un outil en ligne, où les informations sont fournies manuellement pour accéder à une prédiction. Là encore, le nombre d'informations requises doit être limité pour faciliter son utilisation quotidienne. L'idéal est de réserver son usage à des professionnels de santé, qui maîtrisent les données d'entrée et sont capables d'interpréter le résultat obtenu. En cas d'accès complétement ouvert, les patients seraient tentés de l'utiliser sans accompagnement.

La troisième et dernière option est d'implémenter l'outil de prédiction directement dans le système informatique du centre de santé. Cela nécessite de créer une interface, puis de la relier par du code spécifique au SIL[28] ou au dossier médical patient, engendrant un coût non négligeable. Les données (prédicteurs) sont alors renseignées automatiquement dans l'outil, rendant son utilisation plus rapide. Dans ce cas, le nombre de prédicteurs peut être important, jusqu'à plusieurs dizaines, puisque les informations ne sont plus entrées manuellement. Cette méthode est donc particulièrement appropriée pour le suivi longitudinal de biomarqueurs (modèles joints). Comme pour la deuxième option, pour éviter l'effet black box et améliorer l'intelligibilité de chaque résultat rendu, il est important de proposer des explications graphiques, ou les valeurs de Shapley.

[28] SIL : système d'information du laboratoire.

Tableau 3 : Résumé des stratégies d'implémentation en routine de modèles de ML

	Outil		
Critères	**Hors connexion**	**En ligne**	**Intégré au système informatique du centre de santé**
Exemple	Arbre de décision, formule	ABIS, score KTFS[29]	Prescription électronique, imagerie médicale
Facilité d'utilisation	+++	++	+ (Nécessite d'être sur place ou d'utiliser un VPN.[30])
Rapidité	++	+	+++
Explicabilité	+++	++	++
Performance	++	+++	+++
Propension à être diffusable	+++	++	+
Sécurité	+++	++	++
Coût	0	++	+++

Le principal bénéficiaire d'une médecine personnalisée et humaine est bien sûr le patient, et sa relation avec l'équipe soignante. C'est souvent le rôle du néphrologue clinicien de proposer une éducation thérapeutique adaptée à son patient, favoriser l'observance, et être à l'écoute de ses éventuelles questions (sur la maladie, les médicaments, les protocoles de recherche...). Un lien significatif existe entre l'observance et la survie sans rejet. Dans une étude de Villeneuve et al., dès 3 ans post transplantation, la survie sans rejet était significativement plus faible dans le groupe de non-observants que dans le groupe de patients observants (72,7 % contre 88,6 %, P = 0,004) [171]. L'observance est la base d'un suivi thérapeutique pharmacologique efficace, puisqu'elle conditionne le respect des posologies proposées.

Les outils connectés, sites internet et applications, permettraient probablement d'améliorer l'éducation, et la participation *du patient* dans sa prise en charge. Ils permettraient aussi d'alimenter les outils de ML avec des données quasi-instantanées concernant la santé ressentie des patients, voire des constantes physiologiques mesurées à domicile (poids, pression artérielle, glycémie, etc.). A ce jour, il en existe peu dans le domaine de la transplantation. Sêmeia sert principalement à transmettre les informations de suivi, recueillies par le patient transplanté lui-même [172]. Ces informations sont ensuite communiquées au

[29] KTFS : Kidney Transplant Failure Score [170].

[30] VPN : Virtual Private Network, service qui établit une connexion chiffrée et sécurisée entre un ordinateur et un réseau.

médecins grâce à un tableau de bord qui peut envoyer des alertes lors du dépassement des valeurs de référence. L'outil proposerait même des modèles d'IA pour prédire l'observance, le risque de réhospitalisation et la survie du greffon. L'application permet également de communiquer plus facilement avec le patient, et améliorerait l'éducation sur sa pathologie. Néanmoins, la majorité des outils (malheureusement) sont destinés uniquement aux médecins [24,142,173]. En effet, il existe encore des obstacles pour leur utilisation chez les patients transplantés [174] : accès à internet, possession d'un smartphone ou d'une tablette, intérêt des patients pour les contenus en ligne, etc. Dans d'autres pays, c'est l'inverse. Aux Etats-Unis par exemple, le chatbot[31] médical de Google répond déjà aux questions de certains patients de la Mayo Clinic [175], soulevant des questions sur l'utilisation de ces données.

Le patient est donc acteur de sa santé et peut jouer un rôle actif en recherche médicale. Avec son consentement ou sa non-opposition selon les cas, il peut contribuer aux avancées par le partage de ses données, après anonymisation ou pseudonymisation.

[31] Logiciel qui dialogue avec un utilisateur. L'un des plus connu est ChatGPT.

Conclusion

En résumé, dans une première partie de ce travail nous avons mis en avant l'intérêt des estimations d'AUC d'un médicament immunosuppresseur, l'acide mycophénolique, pour proposer de nouvelles posologies chez les enfants et adolescents transplantés rénaux. L'utilisation des doses recommandées sur la base de l'AUC estimée a permis d'augmenter significativement la probabilité d'atteindre les cibles recommandées d'$AUC_{0\text{-}12h}$, et de diminuer significativement la variabilité interindividuelle. Cette étude, comme d'autres qui l'ont précédée, encourage la poursuite de l'adaptation posologique du MMF basée sur l'AUC, y compris en utilisant des outils plus performants comme le ML. Dans un deuxième travail, nous avons utilisé les nombreuses estimations d'AUC d'évérolimus (plusieurs centaines), réalisées à l'aide des modèles de population et estimateurs MAP du site internet ABIS, pour entraîner un algorithme de ML. Nous avons ensuite enrichi la base d'apprentissage avec de très nombreuses simulations (plusieurs milliers) issues d'un modèle de population de la littérature, ce qui a amélioré les performances d'estimation de l'AUC jusqu'à un nombre optimal de simulations d'environ 5 000.

Dans une deuxième partie, nous avons utilisé le ML pour imiter le raisonnement d'un groupe d'experts sur le diagnostic du rejet. L'outil développé a été validé dans de larges bases de données fournies par plusieurs centres de transplantation européens, à l'aide de courbes ROC et de courbes PR. Les classifications obtenues étaient plus proches de la conclusion rendue par les experts locaux, que de l'application stricte de la classification de Banff, qui est pourtant la référence internationale. Enfin, nous avons validé un score pronostique de survie du greffon rénal à 10 ans qu'il est possible de réaliser à la visite systématique de la 1re année post-transplantation. Le score AdGFS est calculable facilement à l'aide d'un arbre de décision utilisant 2-7 variables toujours et partout disponibles.

En conclusion, l'IA en transplantation peut prendre de nombreuses formes, correspondant aux différents algorithmes statistiques et aux différents objectifs. En pharmacologie, les applications sont multiples :

- Sélection de médicaments-candidats, en prédisant leur efficacité/toxicité, avant même leur expérimentation en essais cliniques « in vivo »
- Pharmacocinétique
- Sélection de biomarqueurs dans les domaines -omiques, parmi des centaines de candidats corrélés entre eux
- Recherche d'interactions et d'effets indésirables, autres que ceux découverts lors des essais cliniques (qui n'incluent qu'un échantillon de la population, souvent très sélectionné)
- Pronostic de la maladie, prédiction de l'efficacité d'une stratégie thérapeutique pour un individu en fonction de ses caractéristiques.

Dès lors qu'il existe des prédicteurs efficaces, et des variables à prédire pertinentes pour le suivi du patient, l'IA peut jouer un rôle en médecine de précision. Il convient néanmoins de respecter des étapes importantes pour que son utilisation reste adaptée. La validation externe sur des jeux de données de référence, indépendants et suffisamment grands, est indispensable avant d'envisager une utilisation en routine (à défaut d'essais comparatifs

randomisés). Les critères de performances statistiques doivent être suffisamment exhaustifs pour montrer en toute transparence les limites des modèles.

Références bibliographiques

1. Hariharan S, Johnson CP, Bresnahan BA, Taranto SE, McIntosh MJ, Stablein D. Improved graft survival after renal transplantation in the United States, 1988 to 1996. *N Engl J Med*. 2000;342(9):605-612.

2. Pilch NA, Bowman LJ, Taber DJ. Immunosuppression trends in solid organ transplantation: The future of individualization, monitoring, and management. *Pharmacotherapy*. 2021;41(1):119-131.

3. Masson E. Traitements immunosuppresseurs : mécanismes d'action et utilisation clinique. EM-Consulte. https://www.em-consulte.com/article/1139598/traitements-immunosuppresseurs-mecanismes-d-action. Accessed August 11, 2023.

4. Halloran PF. Immunosuppressive drugs for kidney transplantation. *N Engl J Med*. 2004;351(26):2715-2729.

5. Bergan S, Brunet M, Hesselink DA, et al. Personalized therapy for mycophenolate: consensus report by the international association of therapeutic drug monitoring and clinical toxicology. *Ther Drug Monit*. 2021;43(2):150-200.

6. Prémaud A, Weber LT, Tönshoff B, et al. Population pharmacokinetics of mycophenolic acid in pediatric renal transplant patients using parametric and nonparametric approaches. *Pharmacol Res*. 2011;63(3):216-224.

7. Prémaud A, Debord J, Rousseau A, et al. A double absorption-phase model adequately describes mycophenolic acid plasma profiles in de novo renal transplant recipients given oral mycophenolate mofetil. *Clin Pharmacokinet*. 2005;44(8):837-847.

8. Chen B, Gu Z, Chen H, et al. Establishment of high-performance liquid chromatography and enzyme multiplied immunoassay technology methods for determination of free mycophenolic acid and its application in Chinese liver transplant recipients. *Ther Drug Monit*. 2010;32(5):653-660.

9. Marquet P, Saint-Marcoux F, Prémaud A, et al. Performance of the new mycophenolate assay based on IMPDH enzymatic activity for pharmacokinetic investigations and setup of Bayesian estimators in different populations of allograft recipients. *Ther Drug Monit*. 2009;31(4):443-450.

10. Goirand F, Royer B, Hulin A, Saint-Marcoux F. Évaluation du niveau de preuve du suivi thérapeutique pharmacologique de l'évérolimus. *Therapies*. 2011;66(1):57-61.

11. Kovarik JM, Kahan BD, Kaplan B, et al. Longitudinal assessment of everolimus in de novo renal transplant recipients over the first post-transplant year: pharmacokinetics, exposure-response relationships, and influence on cyclosporine. *Clin Pharmacol Ther*. 2001;69(1):48-56.

12. Miyagi C, Tanaka R, Hirata K, et al. High-Sensitivity and High-Throughput Quantification of Everolimus in Human Whole Blood Using Ultrahigh-Performance Liquid Chromatography Coupled With Tandem Mass Spectrometry. *Ther Drug Monit*. 2022;44(5):633-640.

13. Budde K, Tedesco-Silva H, Pestana JM, et al. Enteric-coated mycophenolate sodium provides higher mycophenolic acid predose levels compared with mycophenolate mofetil: implications for therapeutic drug monitoring. *Ther Drug Monit*. 2007;29(3):381-384.

14. Saint-Marcoux F, Vandierdonck S, Prémaud A, Debord J, Rousseau A, Marquet P. Large scale analysis of routine dose adjustments of mycophenolate mofetil based on global exposure in renal transplant patients. *Ther Drug Monit*. 2011;33(3):285-294.

15. Tett SE, Saint-Marcoux F, Staatz CE, et al. Mycophenolate, clinical pharmacokinetics, formulations, and methods for assessing drug exposure. *Transplant Rev Orlando Fla*. 2011;25(2):47-57.

16. van Gelder T, Hilbrands LB, Vanrenterghem Y, et al. A randomized double-blind, multicenter plasma concentration controlled study of the safety and efficacy of oral mycophenolate mofetil for the prevention of acute rejection after kidney transplantation. *Transplantation*. 1999;68(2):261-266.

17. Knight SR, Morris PJ. Does the evidence support the use of mycophenolate mofetil therapeutic drug monitoring in clinical practice? A systematic review. *Transplantation*. 2008;85(12):1675-1685.

18. Chan L, Hartmann E, Cibrik D, Cooper M, Shaw LM. Optimal everolimus concentration is associated with risk reduction for acute rejection in de novo renal transplant recipients. *Transplantation*. 2010;90(1):31-37.

19. Shipkova M, Hesselink DA, Holt DW, et al. Therapeutic Drug Monitoring of Everolimus: A Consensus Report. *Ther Drug Monit*. 2016;38(2):143-169.

20. Paramètres pharmacocinétiques. https://pharmacomedicale.org/pharmacologie/pharmacocinetique/38-parametres-pharmacocinetiques. Accessed May 24, 2023.

21. Prémaud A, Le Meur Y, Debord J, et al. Maximum a posteriori bayesian estimation of mycophenolic acid pharmacokinetics in renal transplant recipients at different postgrafting periods. *Ther Drug Monit*. 2005;27(3):354-361.

22. Robertsen I, Debord J, Åsberg A, Marquet P, Woillard J-B. A Limited Sampling Strategy to Estimate Exposure of Everolimus in Whole Blood and Peripheral Blood Mononuclear Cells in Renal Transplant Recipients Using Population Pharmacokinetic Modeling and Bayesian Estimators. *Clin Pharmacokinet*. 2018;57(11):1459-1469.

23. Zwart TC, Moes DJAR, van der Boog PJM, et al. Model-Informed Precision Dosing of Everolimus: External Validation in Adult Renal Transplant Recipients. *Clin Pharmacokinet*. 2021;60(2):191-203.

24. ABIS 3.0. https://abis.chu-limoges.fr/login. Accessed May 19, 2023.

25. Riff C, Debord J, Monchaud C, Marquet P, Woillard J-B. Population pharmacokinetic model and Bayesian estimator for 2 tacrolimus formulations in adult liver transplant patients. *Br J Clin Pharmacol*. 2019;85(8):1740-1750.

26. Woillard J-B, Saint-Marcoux F, Monchaud C, Youdarène R, Pouche L, Marquet P. Mycophenolic mofetil optimized pharmacokinetic modelling, and exposure-effect associations in adult heart transplant recipients. *Pharmacol Res*. 2015;99:308-315.

27. Fruit D, Rousseau A, Amrein C, et al. Ciclosporin population pharmacokinetics and Bayesian estimation in thoracic transplant recipients. *Clin Pharmacokinet*. 2013;52(4):277-288.

28. Beaulieu Q, Zhang D, Melki I, et al. Pharmacokinetics of mycophenolic acid and external evaluation of two limited sampling strategies of drug exposure in patients with juvenile systematic lupus erythematosus. *Eur J Clin Pharmacol*. 2022;78(6):1003-1010.

29. REXETRIS : Relations EXposition - Effet à long terme chez le Transplanté Rénal des médicaments ImmunoSuppresseurs. Health Data Hub. https://www.health-data-hub.fr/projets/rexetris-relations-exposition-effet-long-terme-chez-le-transplante-renal-des-medicaments. Accessed May 26, 2023.

30. Hilbrands L, Budde K, Bellini MI, et al. Allograft Function as Endpoint for Clinical Trials in Kidney Transplantation. *Transpl Int*. 2022;35:10139.

31. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Ann Intern Med*. 1999;130(6):461-470.

32. Levey AS, Stevens LA, Schmid CH, et al. A New Equation to Estimate Glomerular Filtration Rate. *Ann Intern Med*. 2009;150(9):604-612.

33. EMA. Clinical investigation of medicinal products to prevent development/slow progression chronic renal insufficiency - Scientific guideline. European Medicines Agency. https://www.ema.europa.eu/en/clinical-investigation-medicinal-products-prevent-development-slow-progression-chronic-renal. Published September 17, 2018. Accessed May 30, 2023.

34. Naesens M, Loupy A, Hilbrands L, et al. Rationale for Surrogate Endpoints and Conditional Marketing Authorization of New Therapies for Kidney Transplantation. *Transpl Int*. 2022;35:10137.

35. Giral M, Taddei C, Nguyen JM, et al. Single-center analysis of 468 first cadaveric kidney allografts with a uniform ATG-CsA sequential therapy. *Clin Transpl*. 1996:257-264.

36. Nicol D, MacDonald AS, Lawen J, Belitsky P. Early prediction of renal allograft loss beyond one year. *Transpl Int Off J Eur Soc Organ Transplant*. 1993;6(3):153-157.

37. Hariharan S, McBride MA, Cherikh WS, Tolleris CB, Bresnahan BA, Johnson CP. Post-transplant renal function in the first year predicts long-term kidney transplant survival. *Kidney Int*. 2002;62(1):311-318.

38. Kaplan B, Schold J, Meier-Kriesche H-U. Poor predictive value of serum creatinine for renal allograft loss. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2003;3(12):1560-1565.

39. Fijter JWD, Mallat MJK, Doxiadis IIN, et al. Increased immunogenicity and cause of graft loss of old donor kidneys. *J Am Soc Nephrol JASN*. 2001;12(7):1538-1546.

40. Terasaki PI. Humoral theory of transplantation. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2003;3(6):665-673.

41. Lionaki S, Panagiotellis K, Iniotaki A, Boletis JN. Incidence and clinical significance of de novo donor specific antibodies after kidney transplantation. *Clin Dev Immunol*. 2013;2013:849835.

42. De Vusser K, Lerut E, Kuypers D, et al. The Predictive Value of Kidney Allograft Baseline Biopsies for Long-Term Graft Survival. *J Am Soc Nephrol JASN*. 2013;24(11):1913-1923.

43. Nasic S, Mölne J, Stegmayr B, Peters B. Histological diagnosis from kidney transplant biopsy can contribute to prediction of graft survival. *Nephrology*. 2022;27(6):528-536.

44. Kasiske BL, Israni AK, Snyder JJ, Skeans MA, Peng Y, Weinhandl ED. A simple tool to predict outcomes after kidney transplant. *Am J Kidney Dis Off J Natl Kidney Found*. 2010;56(5):947-960.

45. Foucher Y, Daguin P, Akl A, et al. A clinical scoring system highly predictive of long-term kidney graft survival. *Kidney Int*. 2010;78(12):1288-1294.

46. Moore J, He X, Shabir S, et al. Development and evaluation of a composite risk score to predict kidney transplant failure. *Am J Kidney Dis Off J Natl Kidney Found*. 2011;57(5):744-751.

47. Schnitzler MA, Lentine KL, Gheorghian A, Axelrod D, Trivedi D, L'Italien G. Renal function following living, standard criteria deceased and expanded criteria deceased donor kidney transplantation: impact on graft failure and death. *Transpl Int Off J Eur Soc Organ Transplant*. 2012;25(2):179-191.

48. Shabir S, Halimi J-M, Cherukuri A, et al. Predicting 5-year risk of kidney transplant failure: a prediction instrument using data available at 1 year posttransplantation. *Am J Kidney Dis Off J Natl Kidney Found*. 2014;63(4):643-651.

49. Loupy A, Aubert O, Orandi BJ, et al. Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. *BMJ*. 2019;366:l4923.

50. Prémaud A, Filloux M, Gatault P, et al. An adjustable predictive score of graft survival in kidney transplant patients and the levels of risk linked to de novo donor-specific anti-HLA antibodies. *PloS One*. 2017;12(7):e0180236.

51. Solez K, Axelsen RA, Benediktsson H, et al. International standardization of criteria for the histologic diagnosis of renal allograft rejection: the Banff working classification of kidney transplant pathology. *Kidney Int*. 1993;44(2):411-422.

52. Racusen LC, Solez K, Colvin RB, et al. The Banff 97 working classification of renal allograft pathology. *Kidney Int*. 1999;55(2):713-723.

53. Racusen LC, Colvin RB, Solez K, et al. Antibody-mediated rejection criteria - an addition to the Banff 97 classification of renal allograft rejection. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2003;3(6):708-714.

54. Racusen LC, Halloran PF, Solez K. Banff 2003 meeting report: new diagnostic insights and standards. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2004;4(10):1562-1566.

55. Solez K, Colvin RB, Racusen LC, et al. Banff '05 Meeting Report: differential diagnosis of chronic allograft injury and elimination of chronic allograft nephropathy ('CAN'). *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2007;7(3):518-526.

56. Solez K, Colvin RB, Racusen LC, et al. Banff 07 classification of renal allograft pathology: updates and future directions. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2008;8(4):753-760.

57. Haas M. The Revised (2013) Banff Classification for Antibody-Mediated Rejection of Renal Allografts: Update, Difficulties, and Future Considerations. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2016;16(5):1352-1357.

58. Loupy A, Haas M, Solez K, et al. The Banff 2015 Kidney Meeting Report: Current Challenges in Rejection Classification and Prospects for Adopting Molecular Pathology. *Am J Transplant*. 2017;17(1):28-41.

59. Loupy A, Haas M, Roufosse C, et al. The Banff 2019 Kidney Meeting Report (I): Updates on and clarification of criteria for T cell- and antibody-mediated rejection. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2020;20(9):2318-2331.

60. Callemeyn J, Ameye H, Lerut E, et al. Revisiting the changes in the Banff classification for antibody-mediated rejection after kidney transplantation. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2021;21(7):2413-2423.

61. Wyld M, Morton RL, Hayen A, Howard K, Webster AC. A systematic review and meta-analysis of utility-based quality of life in chronic kidney disease treatments. *PLoS Med*. 2012;9(9):e1001307.

62. Tong A, Oberbauer R, Bellini MI, et al. Patient-Reported Outcomes as Endpoints in Clinical Trials of Kidney Transplantation Interventions. *Transpl Int Off J Eur Soc Organ Transplant*. 2022;35:10134.

63. Rw E, Dl M, Lp G, et al. The quality of life of patients with end-stage renal disease. *N Engl J Med*. 1985;312(9).

64. Bamoulid J, Staeck O, Halleck F, et al. The need for minimization strategies: current problems of immunosuppression. *Transpl Int Off J Eur Soc Organ Transplant*. 2015;28(8):891-900.

65. Ponticelli C, Passerini P. Gastrointestinal complications in renal transplant recipients. *Transpl Int Off J Eur Soc Organ Transplant*. 2005;18(6):643-650.

66. Metz DK, Holford N, Kausman JY, et al. Optimizing Mycophenolic Acid Exposure in Kidney Transplant Recipients: Time for Target Concentration Intervention. *Transplantation*. 2019;103(10):2012-2030.

67. Razonable RR, Humar A. Cytomegalovirus in solid organ transplant recipients-Guidelines of the American Society of Transplantation Infectious Diseases Community of Practice. *Clin Transplant*. 2019;33(9):e13512.

68. Kant S, Dasgupta A, Bagnasco S, Brennan DC. BK Virus Nephropathy in Kidney Transplantation: A State-of-the-Art Review. *Viruses*. 2022;14(8):1616.

69. Hirsch HH, Randhawa P, AST Infectious Diseases Community of Practice. BK polyomavirus in solid organ transplantation. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2013;13 Suppl 4:179-188.

70. Sawinski D, Goral S. BK virus infection: an update on diagnosis and treatment. *Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc*. 2015;30(2):209-217.

71. Bressollette-Bodin C, Coste-Burel M, Hourmant M, Sebille V, Andre-Garnier E, Imbert-Marcille BM. A prospective longitudinal study of BK virus infection in 104 renal transplant recipients. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2005;5(8):1926-1933.

72. Brennan DC, Agha I, Bohl DL, et al. Incidence of BK with tacrolimus versus cyclosporine and impact of preemptive immunosuppression reduction. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2005;5(3):582-594.

73. Allen UD, Preiksaitis JK, AST Infectious Diseases Community of Practice. Post-transplant lymphoproliferative disorders, Epstein-Barr virus infection, and disease in solid organ transplantation: Guidelines from the American Society of Transplantation Infectious Diseases Community of Practice. *Clin Transplant*. 2019;33(9):e13652.

74. Collins L, Quinn A, Stasko T. Skin Cancer and Immunosuppression. *Dermatol Clin*. 2019;37(1):83-94.

75. Sobiak J, Kamińska J, Głyda M, Duda G, Chrzanowska M. Effect of mycophenolate mofetil on hematological side effects incidence in renal transplant recipients. *Clin Transplant*. 2013;27(4):E407-414.

76. Chapman JR. Chronic calcineurin inhibitor nephrotoxicity-lest we forget. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2011;11(4):693-697.

77. Wijdicks EF, Wiesner RH, Dahlke LJ, Krom RA. FK506-induced neurotoxicity in liver transplantation. *Ann Neurol*. 1994;35(4):498-501.

78. Webster AC, Woodroffe RC, Taylor RS, Chapman JR, Craig JC. Tacrolimus versus ciclosporin as primary immunosuppression for kidney transplant recipients: meta-analysis and meta-regression of randomised trial data. *BMJ*. 2005;331(7520):810.

79. Miller LW. Cardiovascular toxicities of immunosuppressive agents. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2002;2(9):807-818.

80. Liptak P, Ivanyi B. Primer: Histopathology of calcineurin-inhibitor toxicity in renal allografts. *Nat Clin Pract Nephrol*. 2006;2(7):398-404; quiz following 404.

81. Murakami N, Riella LV, Funakoshi T. Risk of metabolic complications in kidney transplantation after conversion to mTOR inhibitor: a systematic review and meta-analysis. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2014;14(10):2317-2327.

82. Almeida CC, Silveira MR, de Araújo VE, et al. Safety of Immunosuppressive Drugs Used as Maintenance Therapy in Kidney Transplantation: A Systematic Review and Meta-Analysis. *Pharmaceuticals*. 2013;6(10):1170-1194.

83. Chapman TM, Perry CM. Everolimus. *Drugs*. 2004;64(8):861-872; discussion 873-874.

84. Wiebe C, Nickerson P. Strategic Use of Epitope Matching to Improve Outcomes. *Transplantation*. 2016;100(10):2048-2052.

85. Vincenti F, Friman S, Scheuermann E, et al. Results of an international, randomized trial comparing glucose metabolism disorders and outcome with cyclosporine versus tacrolimus. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2007;7(6):1506-1514.

86. Ghisdal L, Bouchta NB, Broeders N, et al. Conversion from tacrolimus to cyclosporine A for new-onset diabetes after transplantation: a single-centre experience in renal transplanted patients and review of the literature. *Transpl Int Off J Eur Soc Organ Transplant*. 2008;21(2):146-151.

87. Pascual J, Berger SP, Witzke O, et al. Everolimus with Reduced Calcineurin Inhibitor Exposure in Renal Transplantation. *J Am Soc Nephrol JASN*. 2018;29(7):1979-1991.

88. Vincenti F, Rostaing L, Grinyo J, et al. Belatacept and Long-Term Outcomes in Kidney Transplantation. *N Engl J Med*. 2016;374(4):333-343.

89. El-Charabaty E, Geara AS, Ting C, El-Sayegh S, Azzi J. Belatacept: a new era of immunosuppression? *Expert Rev Clin Immunol*. 2012;8(6):527-536.

90. Best NG, Trull AK, Tan KK, Spiegelhalter DJ, Wreghitt TG, Wallwork J. Blood cyclosporine concentrations and cytomegalovirus infection following heart transplantation. *Transplantation*. 1995;60(7):689-694.

91. Schroeder TJ, Hariharan S, First MR. Relationship between cyclosporine bioavailability and clinical outcome in renal transplant recipients. *Transplant Proc*. 1994;26(5):2787-2790.

92. Starling RC, Hare JM, Hauptman P, et al. Therapeutic drug monitoring for everolimus in heart transplant recipients based on exposure-effect modeling. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2004;4(12):2126-2131.

93. Deppenweiler M, Falkowski S, Saint-Marcoux F, et al. Towards therapeutic drug monitoring of everolimus in cancer? Results of an exploratory study of exposure-effect relationship. *Pharmacol Res*. 2017;121:138-144.

94. Roberts MB, Fishman JA. Immunosuppressive Agents and Infectious Risk in Transplantation: Managing the “Net State of Immunosuppression.” *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2020;73(7):e1302-e1317.

95. Le Meur Y, Büchler M, Thierry A, et al. Individualized mycophenolate mofetil dosing based on drug exposure significantly improves patient outcomes after renal transplantation. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2007;7(11):2496-2503.

96. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232.

97. Friedman J. Stochastic Gradient Boosting. *Comput Stat Data Anal*. 2002;38:367-378.

98. Chen T, Guestrin C. *XGBoost: A Scalable Tree Boosting System*.; 2016:794.

99. Ho TK. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol 1. ; 1995:278-282 vol.1.

100. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. Taylor & Francis; 1984.

101. Poli J-P. Structuration automatique de flux télévisuels. May 2007.

102. Silipo R. Ensemble Models: Bagging & Boosting. *Anal Vidhya*. March 2020. https://medium.com/analytics-vidhya/ensemble-models-bagging-boosting-c33706db0b0b. Accessed June 28, 2023.

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

104. Breiman L. Bagging Predictors. *Mach Learn*. 1996;24(2):123-140.

105. Chowdhury S, Lin Y, Liaw B, Kerby L. Evaluation of Tree Based Regression over Multiple Linear Regression for Non-normally Distributed Data in Battery Performance. November 2021.

106. Is Normalization necessary? · Issue #357 · dmlc/xgboost. GitHub. https://github.com/dmlc/xgboost/issues/357. Accessed July 20, 2023.

107. How do you handle multicollinearity in data science models? Quora. https://www.quora.com/How-do-you-handle-multicollinearity-in-data-science-models. Accessed June 29, 2023.

108. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30.

109. Guthrie NL, Carpenter J, Edwards KL, et al. Emergence of digital biomarkers to predict and modify treatment efficacy: machine learning study. *BMJ Open*. 2019;9(7):e030710.

110. XGBoost. https://kaggle.com/code/dansbecker/xgboost. Accessed June 28, 2023.

111. Tous les modèles de Machine Learning expliqués brièvement. MonCoachData. https://moncoachdata.com/blog/modeles-de-machine-learning-expliques/. Published January 16, 2020. Accessed July 22, 2023.

112. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116-119.

113. Hermsen M, de Bel T, den Boer M, et al. Deep Learning-Based Histopathologic Assessment of Kidney Tissue. *J Am Soc Nephrol JASN*. 2019;30(10):1968-1979.

114. Jayapandian CP, Chen Y, Janowczyk AR, et al. Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains. *Kidney Int*. 2021;99(1):86-101.

115. Bouteldja N, Klinkhammer BM, Bülow RD, et al. Deep Learning-Based Segmentation and Quantification in Experimental Kidney Histopathology. *J Am Soc Nephrol JASN*. 2021;32(1):52-68.

116. Woillard J-B, Labriffe M, Debord J, Marquet P. Mycophenolic Acid Exposure Prediction Using Machine Learning. *Clin Pharmacol Ther*. February 2021.

117. Vaulet T, Divard G, Thaunat O, et al. Data-driven Derivation and Validation of Novel Phenotypes for Acute Kidney Transplant Rejection using Semi-supervised Clustering. *J Am Soc Nephrol JASN*. 2021;32(5):1084-1096.

118. Cippà PE, Sun B, Liu J, Chen L, Naesens M, McMahon AP. Transcriptional trajectories of human kidney injury progression. *JCI Insight*. 2018;3(22).

119. Reeve J, Böhmig GA, Eskandary F, et al. Assessing rejection-related disease in kidney transplant biopsies based on archetypal analysis of molecular phenotypes. *JCI Insight*. 2017;2(12).

120. Aubert A. Machine learning : comment évaluer vos modèles ? Analyses et métriques. Saagie. https://www.saagie.com/fr/blog/machine-learning-comment-evaluer-vos-modeles-analyses-et-metriques/. Published October 12, 2021. Accessed August 1, 2023.

121. Brownlee J. What is the Difference Between Test and Validation Datasets? *MachineLearningMastery.com*. July 2017. https://machinelearningmastery.com/difference-test-validation-datasets/. Accessed August 2, 2023.

122. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge University Press; 2007.

123. Evaluating a machine learning model. Jeremy Jordan. https://www.jeremyjordan.me/evaluating-a-machine-learning-model/. Published July 21, 2017. Accessed August 2, 2023.

124. Precision-recall curves – what are they and how are they used? https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used. Accessed August 2, 2023.

125. Shafi A. How to Learn the Definitions of Precision and Recall (For Good). Medium. https://towardsdatascience.com/precision-and-recall-88a3776c8007. Published April 19, 2022. Accessed August 2, 2023.

126. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*. 2015;10(3):e0118432.

127. Weber LT, Hoecker B, Armstrong VW, Oellerich M, Tönshoff B. Long-term pharmacokinetics of mycophenolic acid in pediatric renal transplant recipients over 3 years posttransplant. *Ther Drug Monit*. 2008;30(5):570-575.

128. van Gelder T. How cyclosporine reduces mycophenolic acid exposure by 40% while other calcineurin inhibitors do not. *Kidney Int*. 2021;100(6):1185-1189.

129. Yabuki H, Matsuda Y, Watanabe T, et al. Plasma mycophenolic acid concentration and the clinical outcome after lung transplantation. *Clin Transplant*. 2020;34(12):e14088.

130. Xiang H, Zhou H, Zhang J, et al. Limited Sampling Strategy for Estimation of Mycophenolic Acid Exposure in Adult Chinese Heart Transplant Recipients. *Front Pharmacol*. 2021;12:652333.

131. Moes DJAR, Press RR, den Hartigh J, van der Straaten T, de Fijter JW, Guchelaar H-J. Population pharmacokinetics and pharmacogenetics of everolimus in renal transplant patients. *Clin Pharmacokinet*. 2012;51(7):467-480.

132. Labriffe M, Woillard J-B, Debord J, Marquet P. Machine learning algorithms to estimate everolimus exposure trained on simulated and patient pharmacokinetic profiles. *CPT Pharmacomet Syst Pharmacol*. May 2022.

133. Woillard J-B, Labriffe M, Debord J, Marquet P. Tacrolimus exposure prediction using machine learning. *Clin Pharmacol Ther*. November 2020.

134. Woillard J-B, Labriffe M, Prémaud A, Marquet P. Estimation of drug exposure by machine learning based on simulations from published pharmacokinetic models: The example of tacrolimus. *Pharmacol Res*. 2021;167:105578.

135. Van Loon E, Gazut S, Yazdani S, et al. Development and validation of a peripheral blood mRNA assay for the assessment of antibody-mediated kidney allograft rejection: A multicentre, prospective study. *EBioMedicine*. 2019;46:463-472.

136. Mertens I, Willems H, Van Loon E, et al. Urinary Protein Biomarker Panel for the Diagnosis of Antibody-Mediated Rejection in Kidney Transplant Recipients. *Kidney Int Rep*. 2020;5(9):1448-1458.

137. Mengel M. Humans versus machines: Who learns faster? Editorial regarding: “Machine learning-supported interpretation of kidney graft elementary lesions in combination with clinical data.” *Am J Transplant*. 2022;22(12):2719-2720.

138. Mannon RB. The Banff schema for antibody-mediated rejection: Lost in translation? *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2019;19(1):9-10.

139. Labriffe M, Woillard J-B, Gwinner W, et al. Machine learning-supported interpretation of kidney graft elementary lesions in combination with clinical data. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. September 2022.

140. Reference Guide to the Banff Classification. *BANFF*. https://banfffoundation.org/central-repository-for-banff-2019-resources-3/. Accessed August 14, 2023.

141. Truchot A, Raynaud M, Kamar N, et al. Machine learning does not outperform traditional statistical modelling for kidney allograft failure prediction. *Kidney Int*. 2023;103(5):936-948.

142. AP-HP F. iTransplant : l’intelligence artificielle au service de la transplantation d’organes pour une médecine de précision. *Fond AP-HP*. April 2019. https://fondationrechercheaphp.fr/itransplant-lintelligence-artificielle-service-de-transplantation-dorganes-medecine-de-precision/. Accessed August 15, 2023.

143. Raynaud M, Aubert O, Divard G, et al. Dynamic prediction of renal survival among deeply phenotyped kidney transplant recipients using artificial intelligence: an observational, international, multicohort study. *Lancet Digit Health*. 2021;3(12):e795-e805.

144. https://transplant-prediction-system.shinyapps.io/AI-DISPO/. Accessed August 28, 2023.

145. About AutoPrognosis 2.0. AutoPrognosis. https://www.autoprognosis.vanderschaar-lab.com/home/about-autoprognosis-2. Accessed August 17, 2023.

146. Becker JU, Seron D, Rabant M, Roufosse C, Naesens M. Evolution of the Definition of Rejection in Kidney Transplantation and Its Use as an Endpoint in Clinical Trials. *Transpl Int*. 2022;35:10141.

147. Hale MD, Nicholls AJ, Bullingham RE, et al. The pharmacokinetic-pharmacodynamic relationship for mycophenolate mofetil in renal transplantation. *Clin Pharmacol Ther*. 1998;64(6):672-683.

148. Rother A, Glander P, Vitt E, et al. Inosine monophosphate dehydrogenase activity in paediatrics: age-related regulation and response to mycophenolic acid. *Eur J Clin Pharmacol*. 2012;68(6):913-922.

149. Strommen AM, Moss MC, Goebel J, Bock M. Donor-specific antibodies in a pediatric kidney transplant population-Prevalence and association with antiproliferative drug dosing. *Pediatr Transplant*. 2019;23(6):e13511.

150. Ponthier L, Marquet P, Moes DJAR, et al. Application of machine learning to predict tacrolimus exposure in liver and kidney transplant patients given the MeltDose formulation. *Eur J Clin Pharmacol*. December 2022.

151. Ponthier L, Ensuque P, Destere A, et al. Optimization of Vancomycin Initial Dose in Term and Preterm Neonates by Machine Learning. *Pharm Res*. 2022;39(10):2497-2506.

152. Fitzsimmons WE, Naesens M. Acute Rejection After Kidney Transplant-An Endpoint Not Predictive of Treatment Effect on Graft Survival. *Transplantation*. June 2023.

153. Klintmalm GB, Vincenti F, Kirk A. Steroid-Responsive Acute Rejection Should Not Be the End Point for Immunosuppressive Trials. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2016;16(11):3077-3078.

154. Kers J, Bülow RD, Klinkhammer BM, et al. Deep learning-based classification of kidney transplant pathology: a retrospective, multicentre, proof-of-concept study. *Lancet Digit Health*. 2022;4(1):e18-e26.

155. Lachenbruch PA, Rosenberg AS, Bonvini E, Cavaillé-Coll MW, Colvin RB. Biomarkers and surrogate endpoints in renal transplantation: present status and considerations for clinical trial design. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2004;4(4):451-457.

156. Modena BD, Kurian SM, Gaber LW, et al. Gene Expression in Biopsies of Acute Rejection and Interstitial Fibrosis/Tubular Atrophy Reveals Highly Shared Mechanisms That Correlate With Worse Long-Term Outcomes. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2016;16(7):1982-1998.

157. Chapman JR. Do protocol transplant biopsies improve kidney transplant outcomes? *Curr Opin Nephrol Hypertens*. 2012;21(6):580-586.

158. _reco366_recommandations_rbp_biopsie_renale_mel.pdf. https://www.has-sante.fr/upload/docs/application/pdf/2022-09/_reco366_recommandations_rbp_biopsie_renale_mel.pdf. Accessed August 29, 2023.

159. Seo J-W, Lee YH, Tae DH, et al. Non-Invasive Diagnosis for Acute Rejection Using Urinary mRNA Signature Reflecting Allograft Status in Kidney Transplantation. *Front Immunol*. 2021;12:656632.

160. Tinel C, Devresse A, Vermorel A, et al. Development and validation of an optimized integrative model using urinary chemokines for noninvasive diagnosis of acute allograft rejection. *Am J Transplant Off J Am Soc Transplant Am Soc Transpl Surg*. 2020;20(12):3462-3476.

161. Karpinski M, Rush D, Jeffery J, Pochinco D, Milley D, Nickerson P. Heightened peripheral blood lymphocyte CD69 expression is neither sensitive nor specific as a noninvasive diagnostic test for renal allograft rejection. *J Am Soc Nephrol JASN*. 2003;14(1):226-233.

162. Masson E. Traitement du rejet humoral aigu. EM-Consulte. https://www.em-consulte.com/article/1307164/traitement-du-rejet-humoral-aigu. Accessed August 11, 2023.

163. Page d'accueil. Health Data Hub. https://www.health-data-hub.fr/. Accessed August 16, 2023.

164. Cara G. France 2030 : l'Inserm et Inria pilotes d'un programme national d'envergure sur la santé numérique. *Salle Presse Inserm*. June 2023. https://presse.inserm.fr/france-2030-linserm-et-inria-pilotes-dun-programme-national-denvergure-sur-la-sante-numerique/67084/. Accessed August 17, 2023.

165. Imrie F, Cebere B, McKinney EF, van der Schaar M. AutoPrognosis 2.0: Democratizing diagnostic and prognostic modeling in healthcare with automated machine learning. *PLOS Digit Health*. 2023;2(6):e0000276.

166. AutoPrognosis - A system for automating the design of predictive modeling pipelines tailored for clinical prognosis. July 2023. https://github.com/vanderschaarlab/autoprognosis. Accessed August 17, 2023.

167. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PloS One*. 2019;14(5):e0213653.

168. Alaa AM, van der Schaar M. Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. *Sci Rep*. 2018;8(1):11242.

169. Rahbar H, Hippe DS, Alaa A, et al. The Value of Patient and Tumor Factors in Predicting Preoperative Breast MRI Outcomes. *Radiol Imaging Cancer*. 2020;2(4):e190099.

170. KTFS score - DIVAT. https://www.divat.fr/biostatistique/recherche-methodo/le-score-ktfs. Accessed August 17, 2023.

171. Villeneuve C, Rousseau A, Rerolle J-P, et al. Adherence profiles in kidney transplant patients: Causes and consequences. *Patient Educ Couns*. 2020;103(1):189-198.

172. Semeia. https://www.semeia.io/produits/nephrowise/. Accessed August 25, 2023.

173. Graft Care – Applications sur Google Play. https://play.google.com/store/apps/details?id=com.cibiltech.ptg.graftCare&hl=fr. Accessed August 16, 2023.

174. Desmet J-M, Dipierdomenico L, Spinogatti N, Treille S, Bardiau F. Quelle est la perception des outils numériques par les patients insuffisants rénaux ? *Néphrologie Thérapeutique*. 2020;16(5):325.

175. Sharing Google's Med-PaLM 2 medical large language model, or LLM. Google Cloud Blog. https://cloud.google.com/blog/topics/healthcare-life-sciences/sharing-google-med-palm-2-medical-large-language-model. Accessed August 17, 2023.

176. Burroughs DL, Lorch G, Guo Y, et al. Noncompartmental pharmacokinetics of three intravenous mycophenolate mofetil concentrations in healthy Standardbred mares. *Vet Dermatol*. 2023;34(3):222-234.

Annexes

Annexe 1. Classification de Banff 2019, d'après Loupy et al. [59]

TABLE 4 Updates of 2019 Banff classification for ABMR, borderline changes, TCMR, and polyomavirus nephropathy. All updates in boldface type[a]

Category 1: Normal biopsy or nonspecific changes
Category 2: Antibody-mediated changes
Active ABMR; all 3 criteria must be met for diagnosis
1. Histologic evidence of acute tissue injury, including 1 or more of the following: • Microvascular inflammation (g > 0 and/or ptc > 0), in the absence of recurrent or de novo glomerulonephritis, although in the presence of acute TCMR, borderline infiltrate, or infection, ptc ≥ 1 alone is not sufficient and g must be ≥ 1 • Intimal or transmural arteritis (v > 0)[b] • Acute thrombotic microangiopathy, in the absence of any other cause • Acute tubular injury, in the absence of any other apparent cause
2. Evidence of current/recent antibody interaction with vascular endothelium, including 1 or more of the following: • Linear C4d staining in peritubular capillaries or medullary vasa recta (C4d2 or C4d3 by IF on frozen sections, or C4d > 0 by IHC on paraffin sections) • At least moderate microvascular inflammation ([g + ptc] ≥2) in the absence of recurrent or de novo glomerulonephritis, although in the presence of acute TCMR, borderline infiltrate, or infection, ptc ≥ 2 alone is not sufficient and g must be ≥1 • Increased expression of gene transcripts/classifiers in the biopsy tissue strongly associated with ABMR, if thoroughly validated
3. Serologic evidence of circulating donor-specific antibodies (DSA to HLA or other antigens). C4d staining or expression of validated transcripts/classifiers as noted above in criterion 2 may substitute for DSA; however thorough DSA testing, including testing for non-HLA antibodies if HLA antibody testing is negative, is strongly advised whenever criteria 1 and 2 are met
Chronic active ABMR; all 3 criteria must be met for diagnosis
1. Morphologic evidence of chronic tissue injury, including 1 or more of the following: Transplant glomerulopathy (cg > 0) if no evidence of chronic TMA or chronic recurrent/de novo glomerulonephritis; includes changes evident by electron microscopy (EM) alone (cg1a) Severe peritubular capillary basement membrane multilayering (ptcml1; requires EM) Arterial intimal fibrosis of new onset, excluding other causes; leukocytes within the sclerotic intima favor chronic ABMR if there is no prior history of TCMR, but are not required
2. Identical to criterion 2 for active ABMR, above
3. Identical to criterion 3 for active ABMR, above, including strong recommendation for DSA testing whenever criteria 1 and 2 are met. **Biopsies meeting criterion 1 but not criterion 2 with current or prior evidence of DSA (posttransplant) may be stated as showing chronic ABMR, however remote DSA should not be considered for diagnosis of chronic active or active ABMR**
Chronic (inactive) ABMR
1. cg > 0 and/or severe ptcml (ptcml1)
2. Absence of criterion 2 of current/recent antibody interaction with the endothelium
3. Prior documented diagnosis of active or chronic active ABMR and/or documented prior evidence of DSA

(Continues)

TABLE 4 (Continued)

C4d staining without evidence of rejection; all 4 features must be present for diagnosis[c]
1. Linear C4d staining in peritubular capillaries (C4d2 or C4d3 by IF on frozen sections, or C4d > 0 by IHC on paraffin sections)
2. Criterion 1 for active or chronic active ABMR not met
3. No molecular evidence for ABMR as in criterion 2 for active and chronic active ABMR
4. No acute or chronic active TCMR, or borderline changes
Category 3: Borderline (Suspicious) for acute TCMR
Foci of tubulitis (t1, t2, or t3) with **mild interstitial inflammation (i1)**, or mild (t1) tubulitis with moderate-severe interstitial inflammation (i2 or i3)
No intimal or transmural arteritis (v = 0)
Category 4: TCMR
Acute TCMR
Grade IA: Interstitial inflammation involving >25% of non-sclerotic cortical parenchyma (i2 or i3) with moderate tubulitis (t2) involving 1 or more tubules, not including tubules that are severely atrophic[d]
Grade IB: Interstitial inflammation involving >25% of non-sclerotic cortical parenchyma (i2 or i3) with severe tubulitis (t3) involving 1 or more tubules, not including tubules that are severely atrophic[d]
Grade IIA: Mild to moderate intimal arteritis (v1), with or without interstitial inflammation and/or tubulitis
Grade IIB: Severe intimal arteritis (v2), with or without interstitial inflammation and/or tubulitis
Grade III: Transmural arteritis and/or arterial fibrinoid necrosis involving medial smooth muscle with accompanying mononuclear cell intimal arteritis (v3), with or without interstitial inflammation and/or tubulitis
Chronic active TCMR[e]
Grade IA: Interstitial inflammation involving >25% of sclerotic cortical parenchyma (i-IFTA2 or i-IFTA3) AND > 25% of total cortical parenchyma (ti2 or ti3) with moderate tubulitis (t2 or **t-IFTA2**) involving 1 or more tubules, not including severely atrophic tubules[d]; other known causes of i-IFTA should be ruled out
Grade IB: Interstitial inflammation involving >25% of sclerotic cortical parenchyma (i-IFTA2 or i-IFTA3) AND > 25% of total cortical parenchyma (ti2 or ti3) with severe tubulitis (t3 or **t-IFTA3**) involving 1 or more tubules, not including severely atrophic tubules[d]; other known causes of i-IFTA should be ruled out
Grade II: Chronic allograft arteriopathy (arterial intimal fibrosis with mononuclear cell inflammation in fibrosis and formation of neointima). This may also be a manifestation of chronic active or chronic ABMR or mixed ABMR/TCMR
Category 5: polyomavirus nephropathy[f]
PVN Class 1 **pvl 1 and ci 0-1**
PVN Class 2 **pvl 1 and ci 2-3 OR** **pvl 2 and ci 0-3 OR** **pvl 3 and ci 0-1**
PVN Class 3 **pvl 3 and ci 2-3**

(Continues)

TABLE 4 (Continued)

[a]Individual Banff lesion scores are defined in Table 5, and it is recommended that these be included in the biopsy report.

[b]It should be noted that these arterial lesions may be indicative of ABMR, TCMR, or mixed ABMR/TCMR. "v" lesions and chronic allograft arteriopathy are only scored in arteries having a continuous media with ≥2 smooth muscle layers.

[c]The clinical significance of these findings may be quite different in grafts exposed to anti–blood group antibodies (ABO-incompatible allografts), where they do not appear to be injurious to the graft and may represent accommodation. However, with anti-HLA antibodies, such lesions may progress to chronic ABMR, and more outcome data are needed.

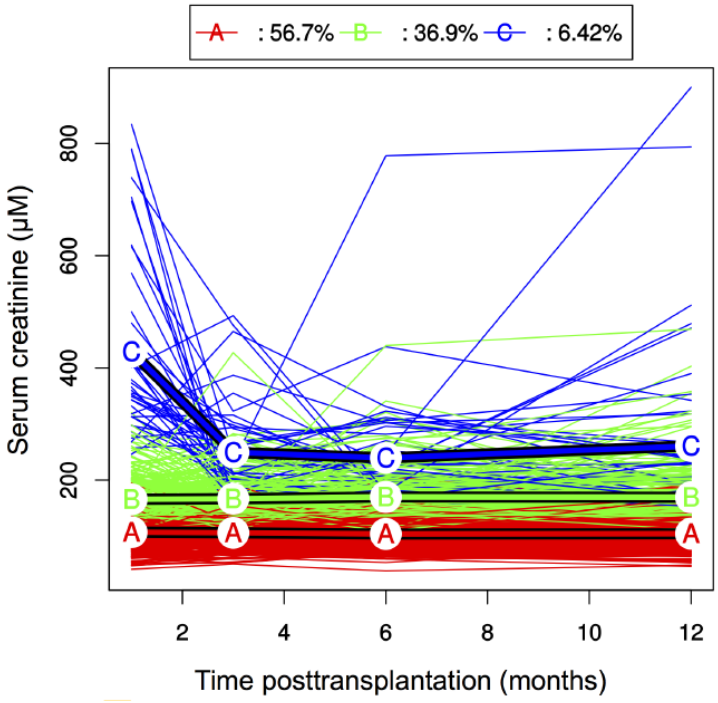
[d]Severely atrophic tubules are defined as having each of 3 features: diameter <25% of unaffected or minimally affected tubules in the same biopsy, an undifferentiated-appearing, cuboidal or flattened epithelium, and pronounced wrinkling and/or thickening of the tubular basement membrane.

[e]It was felt by the majority of Banff 2019 meeting attendees that reporting of chronic active TCMR should be accompanied by a second diagnosis of borderline acute TCMR or acute TCMR (with appropriate grade) when criteria for both diagnoses are met.

[f]pvl scores are defined in Table 5 and in detail in ref. 4. An adequate sample for such scoring should include 2 biopsies cores and contain medulla. PVN can coexist with ABMR or with TCMR grades 2 or 3.

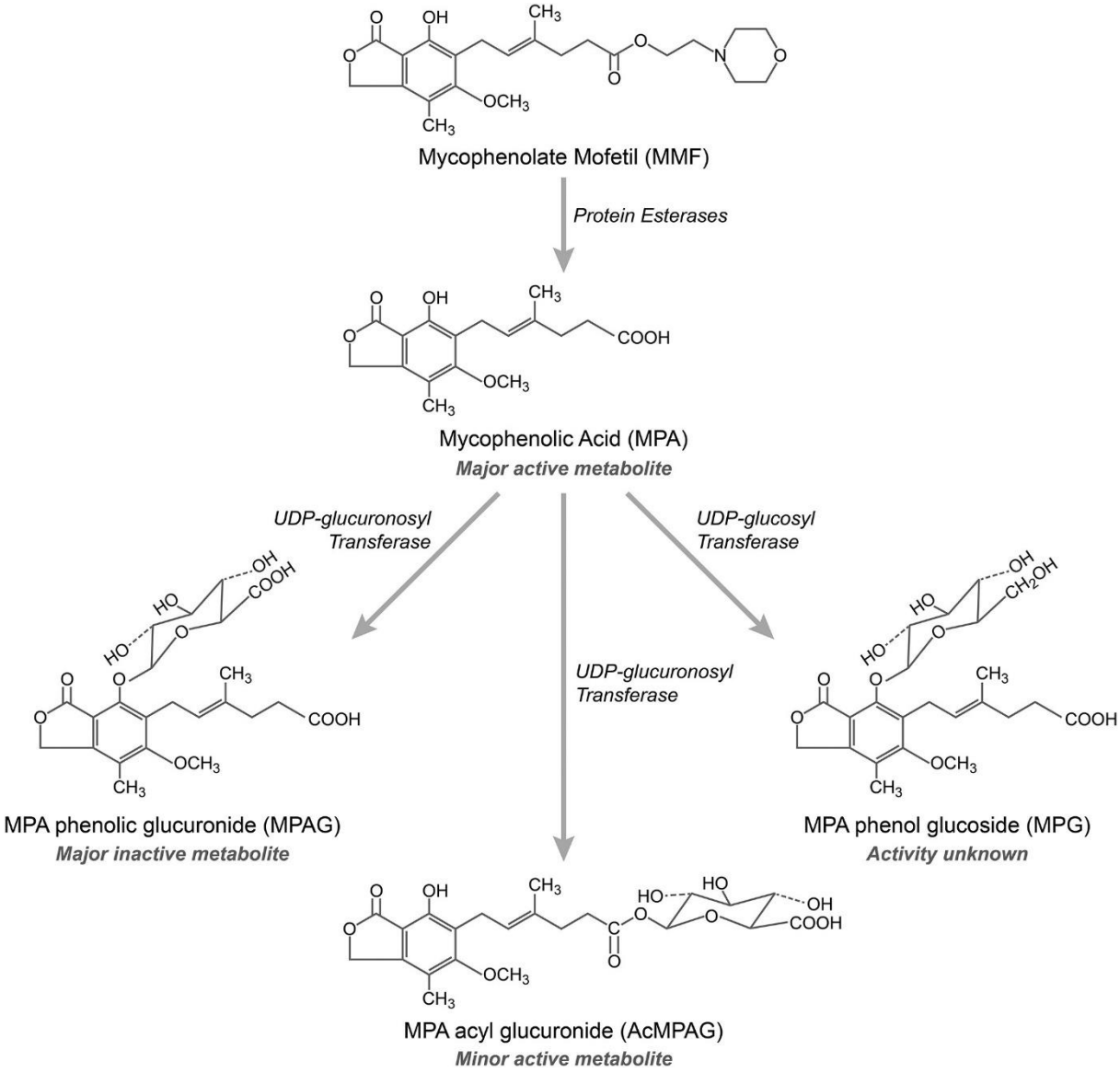
ABMR, rejet médié par les anticorps ; C4d, score de Banff, produit inactif issu de la dégradation catalytique du complément C4 ; ci, score de Banff de fibrose interstitielle ; DSA, anticorps anti-HLA du donneur ; g, score de Banff pour la glomérulite ; i-IFTA, score de Banff d'inflammation dans la fibrose ; i, score de Banff d'inflammation dans l'interstitium ; IHC, immunohistochimie ; ptc, score de Banff de capillarite, présence de cellules dans les capillaires péritubulaires ; ptcml, score de Banff pour la multilamellation de la membrane basale péritubulaire ; pvl, polyomavirus load level, score de Banff gradé de façon semi-quantitative en fonction du pourcentage de tubules cortico-médullaires sièges d'une réplication virale prouvée par immunohistochimie anti-SV40 et/ou par la présence d'inclusions virales dans au moins une cellules épithéliale tubulaire ; t, score de Banff de tubulite ; TCMR, rejet médié par les cellules T ; t-IFTA, score de Banff de tubulite dans la fibrose ; v, score de Banff d'artérite intimale.

Annexe 2. Groupes de profils de créatininémie durant la première année post greffe, d'après Prémaud et al. [50]



Des groupes homogènes de profils de créatininémies ont été identifiés par une méthode basée sur les k-means, adapté pour les données longitudinales (package R *kml*).

Annexe 3. Voies métaboliques du mycophénolate mofétil, d'après Burroughs et al. [176]



Le MMF est hydrolysé par des protéines estérases pour former l'acide mycophénolique (MPA). Le MPA est en outre conjugué à l'uridine diphosphate (UDP)-acide glucuronique pour former soit le principal métabolite inactif glucuronide de l'acide mycophénolique (MPAG), soit le métabolite actif mineur MPA acyl glucuronide (AcMPAG). Une voie métabolique mineure du MPA existe avec l'UDP-glucosyltransférase pour former le MPA phénol glucoside (MPG). L'activité du métabolite MPG est actuellement inconnue.

L’intelligence artificielle au service de la médecine de précision en transplantation

Les travaux présentés ont eu pour objectif de proposer des utilisations appropriées de l’intelligence artificielle (IA) pour la médecine de précision en transplantation. Assurer une prise en charge personnalisée nécessite de combiner de nombreuses informations sur : le patient, le greffon, et les traitements immunosuppresseurs. L’IA offre la possibilité de sélectionner et combiner de nombreuses variables. Dans une première étude, nous avons montré l’apport de l’estimation de l’AUC et des adaptations de posologie par méthode bayésienne (une forme ‘primitive’ d’IA) chez 1 051 transplantés rénaux pédiatriques traités par mycophénolate mofétil. Quand les ajustements de doses proposés étaient suivis, l’intervalle cible d’AUC était plus souvent atteint (p = 0,08 à 0,006) et la variabilité de l’exposition était significativement réduite (p = 0,03 à 0,003). Dans un deuxième travail, nous avons mis au point un algorithme de Machine Learning pour estimer l’$AUC_{0\text{-}12h}$ de l’évérolimus en partant de 508 profils pharmacocinétiques réels, et nous l’avons amélioré en enrichissant progressivement la base d’apprentissage avec des profils simulés (avec un optimal d’environ 5 000 simulations) pour atteindre un écart quadratique moyen (RMSE) de 10,8 µg.h/L en validation externe. Nous avons également mis en évidence les limites d’une telle méthode, avec un surapprentissage à partir de 10 000 simulations se traduisant par une augmentation du RMSE à 12,6 puis 13,7 µg.h/L. Puis, nous avons entraîné un modèle de classification XGBoost sur des diagnostics de rejets humoraux et cellulaires du greffon posés par un groupe d’experts, comme alternative à l’actuelle classification de Banff qui est peu reproductible et ne prend en compte que des données histologiques : des AUC ROC de 0,91 à 0,97 ont été obtenues sur des jeux de données indépendants. Enfin, nous avons validé un score de risque de perte du greffon à long terme, construit à l’aide d’une forêt aléatoire de survie, et utilisant uniquement quelques variables disponibles au premier anniversaire de la transplantation. Le score atteint une AUC ROC = 0,76 et 0,73 à 5 et 10 ans post-transplantation. L’ensemble de ces travaux a donc permis de montrer quelques avantages et limites du Machine Learning pour améliorer la prise en charge médicale des patients transplantés rénaux, comme alternative ou complément des approches statistiques acceptées de plus longue date.

Mots-clés : évérolimus, transplantation rénale, mycophénolate mofétil, acide mycophénolique, ajustement de posologie, pharmacocinétique, apprentissage automatique, simulations

Artificial intelligence at the service of precision medicine in transplantation

The work presented here aimed to propose appropriate uses of artificial intelligence (AI) for precision medicine in transplantation. Ensuring personalized patient care requires combining much information on: the patient, the graft, and immunosuppressive treatments. AI offers the possibility of considering and combining many variables to improve transplant recipients’ long-term outcomes. In a first study, we showed the contribution of mycophenolate mofetil Bayesian (a ‘primitive’ form of AI) AUC estimation and dose adjustment in 1,051 pediatric kidney transplant recipients. When the proposed doses were followed, the AUC target range was more often reached (p = 0.08 to 0.006) and the variability of exposure was significantly reduced (p = 0.03 to 0.003). In a second work, we trained a Machine Learning algorithm to estimate everolimus $AUC_{0\text{-}12h}$ starting with 508 actual pharmacokinetic profiles, and progressively enriched the training set with simulated profiles (up to an optimal number of approx. 5,000 simulations) and reached a root mean square error (RMSE) of 10.8 µg.h/L in an external dataset. We have also highlighted the limits of simulated data, with clear model overfitting when 10,000 simulations or more were used, translating into RMSE increasing to 12.6 and 13.7 µg.h/L. Third, we trained an XGBoost classification model on humoral and cellular rejection diagnoses made by a group of experts, as an alternative to the current Banff classification, which is not very reproducible and only considers histological data. We obtained AUC ROC between 0.91 and 0.97 in independent datasets. Finally, we validated a risk score for long-term graft loss, built using survival random forests, and using only a few variables available at the transplantation first anniversary. The score yields AUC ROC = 0.76 and 0.73, at 5 and 10 years post-transplantation. This piece of work has therefore enabled us to show some advantages and limits of Machine Learning to improve allograft transplant recipients’ medical care, as an alternative or complement to well established statistical approaches.

Keywords: everolimus, kidney transplantation, mycophenolate mofetil, mycophenolic acid, dose adjustment, pharmacokinetic, machine learning, simulations