Ecole Doctorale de Santé Publique

RECOURS AUX MÉTHODES DE POSITIVE AND UNLABELED LEARNING POUR AMÉLIORER LE PHÉNOTYPAGE DES MALADIES DANS LES BASES DE DONNÉES DE SANTÉ

USE OF POSITIVE AND UNLABELED LEARNING METHODS TO IMPROVE DISEASE PHENOTYPING IN HEALTH DATABASES

Etablissement **Université Paris-Sud**

École doctorale **Santé Publique**

Spécialité **santé publique - biostatistiques**

Unité de recherche **Centre d'épidémiologie sur les causes médicales de décès**

Encadrement de la thèse **Grégoire REY (detailResp.pl?resp=38473)**
Co-Directeur **Saïd Karim BOUNEBACHE (detailResp.pl?resp=45225)**

Financement du 01-10-2021 au 30-09-2024

Début de la thèse le **1 octobre 2021**

Date limite de candidature **14 mai 2021**

Mots clés - Keywords

Imputation de données, SNDS, Positive & Unlabeled Learning, Simulation , Analyse de biais

Data Imputation, National Health Data System, Positive & Unlabeled Learning, Simulation , Bias Analysis

Profil et compétences recherchées - Profile and skills required

Le candidat/la candidate devra disposer d'un solide bagage en apprentissage statistique et en manipulation de données complexes, sanctionné par un diplôme de M2 ou d'école d'ingénieur avec spécialisation en mathématiques appliquées, statistiques, ou science des données.

La candidate/le candidat devra témoigner d'une bonne maîtrise du langage de programmation Python.
The candidate must have a solid background in statistical learning and complex data manipulation, leading to an M2 or engineering school diploma with specialization in applied mathematics, statistics, or data science.

The candidate must demonstrate a good level in programming with the Python language.

Description de la problématique de recherche - Project description

Initialement la création du SNDS, qui correspond actuellement à l'alignement des données du SNIIRAM, du PMSI et des causes médicales de décès, a été motivée par la perspective de pouvoir exploiter les parcours de santé des bénéficiaires de l'assurance maladie depuis leur plus jeune âge jusqu'au décès. Cependant, de par leur nature, ces données doivent être utilisées avec précaution notamment lorsqu'on cherche à identifier des bénéficiaires en fonction de pathologies. En effet ces dernières ne sont pas identifiables de façon systématique, aujourd'hui elles sont fournies par un ensemble de règles de décisions à dire d'experts que l'on appelle cartographie. Cette cartographie ne permet d'identifier qu'un sous-ensemble des patients atteints d'une pathologie donnée entrainant potentiellement un ensemble de biais dans les analyses statistiques, voir contraignant ces dernières à restreindre la population cible. Nous avons dans ce cas à faire à des données étiquetées positivement (la pathologie est présente) ou sans étiquettes (nous ne pouvons conclure à la présence ou non de la pathologie), qui nécessitent un traitement particulier. Les méthodes permettant de prendre en compte cette situation s'appelle apprentissage PU (pour positive et unlabeled), et l'objet de cette thèse est d'évaluer l'état de l'art sur la question et de comprendre quel impact auront ces méthodes lors de l'exploitation des données. Cette évaluation se fera à la fois sur des données simulées puis sur un jeu de données réelles permettant de servir de gold standard, issu de la cohorte Constances.
Initially the creation of the National Health Data System, which currently corresponds to the alignment of data from SNIIRAM, PMSI and medical causes of death, was motivated by the prospect of being able to exploit the health pathways of health insurance beneficiaries

since their young age until death. However, by their nature, these data must be used with caution, especially when seeking to identify beneficiaries based on pathologies. Indeed, the latter are not systematically identifiable; today they are provided by a set of decision rules to be told by experts called cartography. This mapping only makes it possible to identify a subset of patients with a given pathology potentially leading to a set of bias in the statistical analyzes, or even forcing them to restrict the target population. In this case, we are dealing with data labeled positively (the pathology is present) or without labels (we cannot conclude on the presence or not of the pathology), which require special treatment. The methods allowing this situation to be taken into account are called PU learning (for positive and unlabeled), and the object of this thesis is to assess the state of the art on the question and to understand what impact these methods will have during data processing. This evaluation will be done both on simulated data and then on a set of real data to serve as a gold standard, from the Constances cohort.

Thématique / Contexte

Le problème de l'apprentissage sur les données du SNDS

Certaines familles de méthodes d'apprentissage machine (ou machine learning) sont utilisées pour la classification supervisée et s'appliquent à des ensembles de données étiquetées [1], [2]. Une donnée est étiquetée si sa classe est connue comme la pathologie pour un patient par exemple. Le Système national des données de santé (SNDS) est le système qui permet de chaîner les données de remboursement de soin de l'assurance maladie, les données hospitalières du programme de médicalisation des systèmes d'information (PMSI), les données de la base des causes médicales de décès produites par le centre d'épidémiologie sur les causes médicales de décès (CépiDc) de l'INSERM ainsi que des données d'autres sources (en quantité marginale). Chacune de ces sources produit ses données pour un objectif bien précis, l'assurance maladie pour rembourser les soins et pour étudier l'évolution de la consommation de soins et des coûts associés. Le PMSI a pour but l'analyse de l'activité hospitalière et permet aux établissements de santé de percevoir le financement de cette activité. Enfin le CépiDc a pour but de produire la statistique nationale des causes médicales de décès. En raison de ces différents objectifs la nature des données varie en fonction des différentes sources.

Fondamentalement les données du SNDS (hors causes de décès) ont des objectifs économiques, les bases de données de l'assurance maladie et de l'ATIH sont construites dans un but de tarification et de pilotage du système de santé. De fait elles ne contiennent pas toujours l'information sur la pathologie du patient sauf si cette dernière implique une prise en charge particulière (par exemple une hospitalisation ou affection de longue durée (ALD)). Par ailleurs, afin de pouvoir réaliser ses études économiques en distinguant le poids des différentes pathologies, la Caisse nationale d'assurance maladie utilisent une cartographie [3] à savoir un ensemble de règles de décision permettant d'imputer des pathologies pour chaque patient à partir des données du SNDS. Ces règles de décisions sont le plus souvent établies à dire d'experts, parfois évaluées ([4], [5] par exemple) et produites en partenariat avec le réseau ReDSiam.

Aujourd'hui, ces données sont de plus en plus utilisées à des fins épidémiologiques. Ceci est possible en raison de leur richesse (variables nombreuses) et leur exhaustivité, puisque chaque utilisation de la carte vitale ou le traitement d'une fiche de soin fait l'objet d'un enregistrement dans cette base.
Néanmoins, il est essentiel de prendre en compte le fait que ces données n'ont pas été constituées initialement pour conduire des études épidémiologiques. Ne pas le faire présente le risque d'introduire des biais qui ont été discutés [6]. Par exemple, avoir une Affection de Longue Durée (ALD) ALD23 « Psychose, trouble grave de la personnalité, arriération mentale » permet d'affirmer la présence d'une pathologie psychiatrique, mais toutes les pathologies psychiatriques n'entrainent pas une ALD23. En effet, les patients peuvent ne pas être sous ALD ou bien avoir une ALD pour une autre cause rendant peu utile la déclaration de cette nouvelle ALD. Dès lors, il peut s'avérer délicat de chercher à constituer un ensemble d'apprentissage pour distinguer les patients ayant une pathologie psychiatrique des autres. De même, les bases de données des hôpitaux sont construites pour suivre la prise en charge des patients ; la réutilisation des données de soins pose donc des défis informationnels pour leur usage dans des études épidémiologiques. Le repérage à partir des données structurées de certains groupes de patients repose sur l'expertise des Départements d'Information Médicale (DIM) des hôpitaux avec un choix de codes de la Classification Internationale des Maladies (CIM) et/ou d'actes pouvant reposer sur des études de validations [7], [8].

Finalement, une difficulté récurrente est que lorsque sont utilisées les données du SNDS pour étudier les facteurs d'incidence de la pathologie A, il n'existe pas de variables permettant d'étiqueter des témoins. Les cas correspondent alors aux étiquetages positifs et les autres sont tout simplement non étiquetés (unlabeled).

Dans ce cadre, que l'on appellera Positive & Unlabeled (PU) learning, on souhaite construire un modèle de classification alors qu'on ne dispose que d'un ensemble d'éléments ayant un statut A connu. François Denis a démontré que sous certaines conditions (quantité d'information suffisante et hypothèses sur les distributions sous-jacentes), il est possible de réaliser des apprentissages sur des données positives [9], [10]. Initialement, les méthodes de PU learning reposent sur des algorithmes en deux étapes réalisées de manières itératives. La première étape consiste à identifier et extraire des exemples négatifs probables (RN) de l'ensemble des données non étiquetées (U). La seconde étape consiste à entrainer un modèle de classification à partir de l'ensemble des données étiquetées (P) et de l'ensemble RN. Ces deux étapes sont répétées jusqu'à la convergence vers une solution stable [11]. Mais la littérature sur le sujet s'est

beaucoup développée ces dernières années.

Un cas d'usage : la cohorte Constances comme base de validation.

L'une des raisons pour laquelle ces méthodes se sont peu développées est qu'il existe peu de bases qui puissent servir d'ensemble de validation. La cohorte Constances est une cohorte épidémiologique à vocation généraliste. Elle est constituée d'un échantillon représentatif de 200 000 personnes âgées de 18 à 69 ans à l'inclusion. L'avantage de cette cohorte est qu'elle tire au sort ses éléments parmi les assurés de l'assurance maladie du régime général de la sécurité sociale de 17 départements. Suivant le protocole, les membres de la cohorte doivent passer un examen médical tous les 5 ans et répondre à un questionnaire tous les ans. La cohorte Constances contient des données du SNDS alignées avec des données fiables issues d'investigation clinique permettant de mesurer l'écart obtenu entre un algorithme d'identification de pathologie à partir du SNDS et un gold standard. C'est d'ailleurs ce qui a été entrepris sur le diabète dans [12], où il a été montré qu'il est possible d'augmenter significativement la sensibilité des algorithmes de ciblage du diabètes (sensibilité inférieure à 80% en utilisant uniquement les ALD, et de plus de 93% en utilisant un algorithme plus complexe). La cohorte Constances peut donc être considérée comme une source très riche pour travailler sur un ensemble de pathologies en utilisant la cartographie de la Cnam pour constituer des jeux de données PU.

Initialement les méthodes de PU learning sont principalement utilisées dans le domaine de la classification de textes [13]. D'autres approches ont été développées pour limiter la taille des ensembles d'apprentissage comme le recours à l'active learning [14], [15] et le reinforcement learning [16].
Hypothèses de recherche
Nous faisons l'hypothèse que les méthodes de PU learning et l'intégration de connaissances externes (connaissances sur la structure des données) pourraient permettre de construire des modèles de classification des maladies guidés par les données.

Objectifs

Ce travail de recherche devra permettre de comprendre la place relative des approches de PU Learning dans le problème de phénotypage des maladies pour des ensembles non étiquetées. Les algorithmes de PU learning en deux étapes permettent de construire des modèles à partir d'un ensemble d'individus étiquetés positivement, c'est une approche reposant sur des modèles de classification. La recherche de variables à ancre est une approche reposant sur la recherche de variables d'intérêt permise à la fois par l'exploration de la structure des données et le choix des experts. Enfin les méthodes d'apprentissage profond de par leur grande capacité de représentation sont capables d'apprendre la structure géométrique des données à condition d'avoir des données en grande quantité. Ces trois approches ne sont pas totalement en compétition puisque leur performance sera très dépendante de la prévalence réelle des pathologies. Il sera envisagé d'avoir recours à un agrégateur pour les différents systèmes de classification obtenus comme le super learner [24] pour les pathologies à plus faible prévalence.

La thèse s'axera autour des trois objectifs suivants :

• On cherchera à produire un outil permettant de simuler des données complexes en regardant du côté des réseaux génératifs antagonistes afin de pouvoir générer des situations complexes. Ces outils nous permettront de pouvoir comprendre ce qui se passe dans le PU-learning du fait de pouvoir contrôler des scénarios ce que ne nous permet pas l'utilisation de données réelles.

• La recension et l'évaluation des méthodes de PU learning sur des données simulées, et notamment mesurer l'impact des hypothèses sur la validité de ces approches. Le générateur de données pourra être basé sur des architectures GAN afin de pouvoir créer des modèles plus complexes. Tout comme la règle de Rubin permet de fournir un estimateur moins biaisé ainsi que de prendre en compte l'incertitude, dans une analyse statistique, due à la présence de données manquantes, le second objectif consistera en la production de règles, voire de méthodes permettant d'intégrer cette incertitude en prenant en compte les nouveaux attributs obtenus par le PU learning.

• On appliquera ces méthodes sur les données de la cohortes Constances, et plus précisément dans un premier temps on se basera sur ce qui a déjà été entrepris sur le diabète [12]. Par ailleurs on utilisera la cartographie de la Cnam pour créer d'autres jeux de données PU, un travail sera fait pour sélectionner d'autres pathologies à inclure dans l'évaluation.

Méthode

Deux concepts sont sous-jacents à l'exploration de ce sujet : les algorithmes de PU learning « iteratives » et l'exploration de la structure des données par le moyen de variables à ancre (anchor variables).
Les méthodes de PU learning en deux étapes
La première étape consiste à extraire un ensemble d'éléments probablement négatifs (RN) de l'ensemble non étiqueté (U). Plusieurs méthodes ont été développées pour résoudre ce problème : la méthode de Rocchio, l'utilisation des méthodes de classification naïve bayésienne (NB), la spy technique, l'utilisation de formes normales disjonctives (1-DNF) [17].
L'approche par NB consiste à estimer une distribution binomiale de probabilité à partir des ensembles P et U et de considérer les individus RN comme ceux ayant une probabilité forte de ne pas appartenir à P. La spy technique est similaire à l'approche par NB sauf qu'une partie de l'ensemble P est extraite (notée SP) et est réunie à U avant d'appliquer le classificateur NB. Le modèle est alors entrainé sur P et

U∪SP et les éléments de SP permettent de déterminer le seuil à partir duquel un élément peut raisonnablement être considéré comme appartenant à l'ensemble RN. La méthode Rocchio consiste à définir un vecteur de classes pour chaque élément de l'ensemble et à déterminer les distances entre les éléments à l'aide des mesures cosinus entre les vecteurs de classes. Cela permet donc d'identifier des éléments appartenant à l'ensemble RN en les considérant comme éloignés des vecteurs de l'ensemble P. Le 1-DNF consiste à déterminer les variables qui apparaissent plus fréquemment dans l'ensemble P que dans l'ensemble U. Une fois ces variables identifiées, on considère les éléments RN de U comme ceux pour lesquels aucune de ces variables n'est présente.
La seconde étape consiste à entrainer le modèle de classification à partir des ensembles P (ou P-S) et RN. Les algorithmes utilisés peuvent être des SVM (Support Vector Machine) ou des NB. Pour certaines méthodes, les étapes 1 et 2 appliquées de manière itérative comme dans l'algorithme PEBL [18] qui utilise une combinaison 1-DNF et SVM.
Approche des variables à ancre (anchor variables)
David Sondag et son équipe ont développé le concept de variable à ancre ou anchor variable. La variable X_i est une ancre positive pour Y si X_i est indépendant de X_j conditionnellement à Y et que P(Y=1 | X_i=1)=1. Connaître X_i permet donc de déterminer la positivité de Y, en revanche si X_i=0, cela ne signifie pas que Y soit négatif. L'idée de cette approche vient de la démarche du PU learning bien que ces méthodes ne s'inscrivent pas dans les algorithmes en deux étapes présentées ci-dessus. Elle consiste à accompagner les experts dans la recherche de variables à ancre en sélectionnant les candidats potentiels. Sontag et al. ont ainsi développé un outil d'élicitation permettant d'accompagner les experts dans le choix de variables à ancre [19], [20]. Ces variables sont alors utilisées pour étiqueter un certain nombre d'éléments non étiquetées, ou bien pour repondérer des unités statistiques lors de l'apprentissage.

Méthode issue de l'apprentissage profond

Les modèles de l'apprentissage profond (Deep Learning) sont connus pour leur pouvoir de représentation et leur capacité à détecter la structure des données. Il était donc tout à fait naturel d'adapter ces méthodes au PU learning. Parmi les architectures connues, les réseaux génératifs par antagonismes [21] sont devenus un outil puissant pour reproduire la loi de probabilité des données. Consistant en la mise en compétition entre un modèle de discrimination (donné par une architecture de réseau de neurones) dont le but est de détecter les vraies données et les données simulées et un modèle génératif (image d'une loi de probabilité par un réseau de neurones), [22] utilise ce type d'architecture pour reconstituer la distribution des étiquettes négatives. [23] tente plutôt de déterminer la structure géométrique qui discriminera au mieux les deux groupes en ajoutant une pénalisation au risque empirique.

Résultats attendus - Expected results

Les contributions attendues de ces travaux de thèse se situent à différents niveaux :

• En épidémiologie, puisque ce travail fournira une méthodologie permettant une exploitation « plus juste » des données du SNDS.
• En biostatistique puisqu'on cherchera à fournir de nouvelles méthodes que l'on peut assimiler à de l'imputation de données. Par ailleurs nous chercherons à fournir de nouvelles méthodes de simulation non paramétrique essayant de reproduire des caractéristiques complexes des données.

Précisions sur l'encadrement - Details on the thesis supervision

La thèse sera co-encadrée par Ismaïl Ahmed de l'équipe Biostatistique en grande dimension du CESP. Des points très réguliers seront faits avec les encadrants, et au minimum une réunion de suivi mensuelle sera organisée avec l'ensemble de l'équipe encadrante. Ils permettront notamment de suivre l'état d'avancement des travaux, de leur valorisation et des formations suivies.

Conditions scientifiques matérielles et financières du projet de recherche

La Cohorte Constances utilise actuellement le CASD (le Centre d'Accès Sécurisé aux Données) pour diffuser ces données. Ce service a mis en place une bulle sécurisée avec des cartes graphiques NVIDIA TESLA T4, permettant le déploiement d'un large panel d'algorithmes d'apprentissage machine incluant l'apprentissage profond.

Objectifs de valorisation des travaux de recherche du doctorant : diffusion, publication et confidentialité, droit à la propriété intellectuelle,...

Dans un premier temps nous projetons de faire une revue des méthodes de PU learning que nous évaluerons sur les données de Constances avec le diabète ce qui fera l'objet d'une première publication en première année de thèse, ici nous ne regarderons que le pouvoir prédictif du diabète.
Pour une compréhension plus profonde de ces méthodes, une étude sera faite en utilisant des simulations, afin de mettre à l'épreuve les hypothèses sous-jacentes. Nous utiliserons par ailleurs des méthodes de simulations utilisant des architectures GAN permettant de simuler des données réaliste à la fois pour la création de scenario mais aussi pour l'imputation de pathologies. Nous développerons des règles méthodologiques pour prendre en compte les outcomes fournis par ces algorithmes (imputation, scores, etc.), cette seconde partie fera l'objet d'un second article pour la deuxième année de thèse.
Enfin nous passerons à l'échelle en appliquant ces méthodes sur d'autres pathologies de la cartographie de la Cnam afin de voir

l'applicabilité réelle de ces méthodes notamment en lien avec la prévalence des maladies et de porter un regard critique sur ces outils. Cette dernière année de thèse se conclura avec la rédaction du troisièmes article et du manuscrit de thèse.

Références bibliographiques

[1] J. Iavindrasana, G. Cohen, et A. Depeursinge, « Clinical data mining: a review », Yearb Med, 2009.
[2] S. H. Liao, P. H. Chu, et P. Y. Hsiao, « Data mining techniques and applications - A decade review from 2000 to 2011 », Expert Systems with Applications, vol. 39, no 12. p. 11303-11311, 2012, doi: 10.1016/j.eswa.2012.02.063.
[3] « Méthodologie de repérage des pathologies et de répartition des dépenses par pathologie », 2017.
[4] K. Goueslard, J. Cottenet, E. Benzenine, P. Tubert-Bitter, et C. Quantin, « Validation study: evaluation of the metrological quality of French hospital data for perinatal algorithms », BMJ Open, vol. 10, no 5, p. e035218, mai 2020, doi: 10.1136/bmjopen-2019-035218.
[5] B.-L. Pauline et al., « Validation d'un algorithme complexe d'identification de poussées dans la sclérose en plaque (SEP) à partir du Système National des Données de Santé (SNDS) », Rev. Neurol. (Paris), vol. 176, p. S81-S82, sept. 2020, doi: 10.1016/j.neurol.2020.01.243.
[6] P. Tuppin et al., « Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France », Rev. DÉpidémiologie Santé Publique, juill. 2017, doi: 10.1016/j.respe.2017.05.004.
[7] L. Tamariz, T. Harkins, et V. Nair, « A systematic review of validated methods for identifying ventricular arrhythmias using administrative and claims data », Pharmacoepidemiol. Drug Saf., vol. 21, p. 148-153, janv. 2012, doi: 10.1002/pds.2340.
[8] C. P. Chung, P. Rohan, S. Krishnaswami, et M. L. McPheeters, « A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data », Vaccine, vol. 31, p. K41-K61, déc. 2013, doi: 10.1016/j.vaccine.2013.03.075.
[9] F. Denis, « PAC Learning from Positive Statistical queries », Algorithmic Learn. Theory, p. 112-126, 1998, doi: 10.1007/3-540-49730-7_9.
[10] F. Denis, R. Gilleron, et F. Letouzey, « Learning from positive and unlabeled examples », Theor. Comput. Sci., vol. 348, no 1, p. 70-83, déc. 2005, doi: 10.1016/j.tcs.2005.09.007.
[11] Y. Chen, « Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets », 2008.
[12] S. Fuentes et al., « Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort », Int. J. Public Health, vol. 64, no 3, p. 441-450, avr. 2019, doi: 10.1007/s00038-018-1186-3.
[13] B. Liu, Y. Dai, X. Li, W. S. Lee, et P. S. Yu, « Building text classifiers using positive and unlabeled examples », in Third IEEE International Conference on Data Mining, p. 179-186, doi: 10.1109/ICDM.2003.1250918.
[14] J. Pathak, A. N. Kho, et J. C. Denny, « Electronic health records-driven phenotyping: challenges, recent advances, and perspectives », J. Am. Med. Inform. Assoc., vol. 20, no e2, p. e206-e211, déc. 2013, doi: 10.1136/amiajnl-2013-002428.
[15] Y. Chen et al., « Applying active learning to high-throughput phenotyping algorithms for electronic health records data », J. Am. Med. Inform. Assoc., vol. 20, no e2, p. e253-e259, déc. 2013, doi: 10.1136/amiajnl-2013-001945.
[16] J. Rennie et A. K. McCallum, « Using reinforcement learning to spider the Web efficiently », Proc. ICML-99 16th Int. Conf. Mach. Learn., p. 335-343, 1999.
[17] B. Liu, Y. Dai, X. Li, W. S. Lee, et P. Yu, « Building Text Classifiers Using Positive and Unlabeled Examples », 2003.
[18] H. Yu, J. Han, et K. Chen-Chuan Chang, « PEBL: Web Page Classification without Negative Examples ».
[19] Y. Halpern, Y. Choi, S. Horng, et D. Sontag, « Using Anchors to Estimate Clinical State without Labeled Data. », AMIA Annu. Symp. Proc. AMIA Symp., vol. 2014, p. 606-15, 2014.
[20] Y. Halpern, S. Horng, Y. Choi, et D. Sontag, « Electronic medical record phenotyping using the anchor and learn framework », J. Am. Med. Inform. Assoc., vol. 23, no 4, p. 731-740, juill. 2016, doi: 10.1093/jamia/ocw011.
[21] I. J. Goodfellow et al., « Generative Adversarial Networks », ArXiv14062661 Cs Stat, juin 2014, Consulté le: mars 19, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/1406.2661.
[22] M. Hou, B. Chaib-draa, C. Li, et Q. Zhao, « Generative Adversarial Positive-Unlabelled Learning », in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, juill. 2018, p. 2255-2261, doi: 10.24963/ijcai.2018/312.
[23] « Deep-PUMR: Deep Positive and Unlabeled Learning with Manifold Regularization », springerprofessional.de. https://www.springerprofessional.de/en/deep-pumr-deep-positive-and-unlabeled-learning-with-manifold-reg/16310016 (consulté le mars 18, 2021).
[24] M. J. van der Laan, E. C. Polley, et A. E. Hubbard, « Super learner », Stat. Appl. Genet. Mol. Biol., vol. 6, p. Article25, 2007, doi: 10.2202/1544-6115.1309.

Dernière mise à jour le 17 avril 2021