CO10.6 - Imputation des données manquantes par un méta-algorithme (metaCART): étude de simulation

I. El Badisy, C. Nejjari, A. Naim, K. El Rhaz, M. Khalis, R. Giorgi

recueillies par Météo France, Google et différents services de santé publique.

Résultats : Nous avons étudié les propriétés théoriques de l'estimateur issu de cette régression et apportons des garanties théoriques concernant cet estimateur. Sur les données COVID, nous obtenons une meilleure performance de notre algorithme comparativement aux méthodes concurrentes d'équations différentielles contrôlées neuronales [5], GRU [2] et de combinaisons d'estimateurs issus de modèles d'apprentissage statistique.

Conclusion : Les signatures sont un outil performant pour analyser des données de santé.

Mots clés : Apprentissage statistique , Epidémiologie , Signatures , Séries temporelles , Systèmes dynamiques

Déclaration de liens d'intérêts : Les auteurs n'ont pas précisé leurs éventuels liens d'intérêts.

Références

[1] Kuo-Tsai Chen. "Integration of paths−A faithful representation of paths by non- commutative formal power series". In: Transactions of the American Mathematical Society 89.2 (1958), pp. 395−407.
[2] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: NIPS 2014 Workshop on Deep Learning, December 2014. 2014.
[3] Adeline Fermanian. "Embedding and learning with signatures". In: Computational Statistics & Data Analysis 157 (2021), p. 107148.
[4] P. Kidger, et al. "Deep signature transforms". In: Advances in Neural Informa- tion Processing Systems 32 (2019).
[5] P. Kidger, et al. "Neural controlled differential equations for irregular time se- ries". In: Advances in Neural Information Processing Systems 33 (2020), pp. 6696-707.
[6] T. Lyons, et al. Differential equations driven by rough paths. Springer, 2007.
[7] T. Lyons, et al. "Signature Methods in Machine Learning". In: arXiv preprint arXiv:2206.14674 (2022).
[8] J. Paireau, et al. "An ensemble model based on early predictors to forecast COVID-19 health care demand in France". In: Proceedings of the National Academy of Sciences 119.18 (2022), e2103302119.

https://doi.org/10.1016/j.respe.2023.101630

CO10.5 - Prédiction du risque de décès au décours d'une circulation extra-corporelle avec oxygénateur à membranes veino-artérielle (ECMO-VA), grâce à des méthodes d'intelligence artificielle

K. Sallah[1,2,*], A. Balzano[1,2], P. Eloy[1,2], J. Do-Vale[3], C. De Tymowski[3], S. Provenchere[3]

[1] *Inserm, CIC-EC 1425, Paris, France*
[2] *AP-HP Nord-Université Paris Cité, HUPNVS, Clinical Research, Biostatistics and Epidemiology Department, Paris, France*
[3] *AP-HP DMU PARABOL, Department of Anaesthesiology and Surgical Intensive Care, Paris, France*
[*] *Auteur correspondant*
E-mail address: kankoe.sallah@aphp.fr

Introduction : La circulation extra-corporelle par membranes d'oxygénation de type veino-artérielle (ECMO-VA) est la technique d'assistance pour la prise en charge des chocs cardiogéniques. Cette technique est invasive et de disponibilité limitée. Une meilleure connaissance des facteurs prédictifs de mortalité post-ECMO-VA permettrait de mieux sélectionner les patients pouvant utilement bénéficier de cette thérapeutique. Notre objectif était de comparer les prédictions de mortalité à 90 jours obtenues à l'aide des méthodes d'intelligence artificielle à celles obtenues en utilisant des scores classiques.

Méthodes : Les données de soins de 1356 patients adultes ayant bénéficié d'une ECMO-VA entre août 2017 et novembre 2021 ont été extraites de l'Entrepôt de données de santé (EDS) de l'Assistance publique−Hôpitaux de Paris. Age, sexe, poids, taille, antécédents, ainsi que diagnostics, actes médicaux, données de biologie pendant le séjour et statuts vitaux à J90 ont été recueillis soit à partir des tables structurées, soit à l'aide d'Expressions régulières (Regex). Les scores classiques SAVE, ENCOURAGE et REMEMBER et autres scores de morbidité utilisés en pratique courante par les cliniciens, ont été reconstruits sur la base des informations contenues dans les dossiers médicaux. Les variables présentant plus de 25 % de données manquantes ont été exclues. Trois modèles prédictifs ont été élaborés après split 80/20 sur un fichier initial contenant 33 variables explicatives potentielles, retenues par les cliniciens: régression logistique avec sélection de variables (RL), « eXtreme Gradient Boosting » (XGBoost) et « Deep Neural Network » (DNN). Une imputation simple par la médiane a été utilisée pour les données manquantes. L'apprentissage a été renforcée par cross-validation. Les performances des modèles et celles des scores classiques ont été comparées grâce à l'aire sous la courbe ROC (AUC) et test de Delong. La validité des Regex a été évaluée sur un jeu de données restreint.

Résultats : Les méthodes Regex utilisées dans l'extraction des données renvoyaient un taux moyen de bon classement à 82 %. RL, XGBoost et DNN montrent des performances statistiquement équivalentes pour la prédiction du décès à 90 jours, AUC respectifs avec IC à 95% : 0,83[0,78-0,88] ; 0,81[0,76-0,86] et 0,82[0,77-0,87] sur le jeu de données test tandis que les scores classiques affichent des performances plus modestes, le meilleur étant 0,7[0,67-0,73] pour le SAVE score, p-value<0,001, test de Delong. Les analyses de sensibilité introduisant de nouvelles méthodes d'imputation (GAIN imputation, MICE imputation) ainsi que d'autres méthodes de réduction de dimension (ACP, sélection sur Chi2) n'ont pas amélioré les résultats.

Conclusion : Dans notre étude, les modèles proposés se sont révélés plus performants que les scores classiques utilisés pour la prédiction du risque de mortalité au sein d'une cohorte de 1356 patients. Ces résultats encourageants méritent d'être reproduits sur des jeux de données plus vastes et encore plus complets car quelques variables ont été exclues de notre étude en raison d'un grand nombre de données manquantes.

Mots clés : Intelligence artificielle , ECMO-VA , Régression logistique , eXtreme Gradient Boosting , Deep Neural Network

Déclaration de liens d'intérêts : Les auteurs n'ont pas précisé leurs éventuels liens d'intérêts.

https://doi.org/10.1016/j.respe.2023.101631

CO10.6 - Imputation des données manquantes par un méta-algorithme (metaCART): étude de simulation

I. El Badisy[1,2,*], C. Nejjari[1,3], A. Naim[1], K. El Rhaz[3], M. Khalis[1], R. Giorgi[2,4]

[1] *Université Mohammed VI des sciences de la santé (UM6SS), Centre Mohammed VI pour la recherche et l'innovation, Casablanca, Maroc*
[2] *Aix Marseille Université, Inserm, IRD, Sesstim, Sciences économiques & sociales de la santé & traitement de l'information médicale, ISSPAM, Marseille, France*
[3] *Université Sidi Mohamed Ben Abdellah, Laboratoire d'épidémiologie de recherche clinique et de santé communautaire, Fès, Maroc*

[4] Aix Marseille Université, APHM, Inserm, IRD, Hôpital de la Timone, BioSTIC, Biostatistique et technologies de l'information et de la communication, Marseille, France
*Auteur correspondant
E-mail address: elbadisyimad@gmail.com

Introduction : L'imputation multiple est rarement appliquée dans les analyses basées sur l'apprentissage automatique, où la majorité des algorithmes prennent comme input un jeu de données unique. Alors que pour les analyses axées sur l'inférence statistique (estimations des effets des facteurs, significativité statistique), l'imputation multiple reste le gold standard. Par ailleurs, les méthodes basées sur les arbres sont plus robustes aux valeurs aberrantes et peuvent capturer plus efficacement les interactions non linéaires entre les variables. L'objectif de notre de travail consiste à évaluer les performances d'un nouvel algorithme d'imputation inspiré du méta-apprentissage à d'autres algorithmes d'imputation simples et multiples.

Méthodes : Dans cette étude, nous comparons les performances de plusieurs méthodes d'imputation : "mice", "miceRF", "miceCART", "KNN", "CART", "missForest" et "missCforest", à une nouvelle méthode méthode d'imputation : "metaCART". Il s'agit d'un méta-algorithme qui prend comme imputeur de base "KNN", "missCforest" et "missForest". Dans une étude de simulation, nous avons généré 600 jeux de données complets de 1000 patients et ensuite introduit une proportion de 30 % de données manquantes aux covariables sous le mécanisme MCAR. Le modèle de Cox a été utilisé comme modèle substantif pour la génération des données et l'évaluation post-imputation. La performance d'imputation des méthodes a été évaluée par trois paramètres de simulation, le bias, bias relatif et MSE des HRs.

Résultats : A l'aide de l'étude de simulation, nous démontrons que notre approche peut donner lieu à des imputations plus plausibles et donc à des inférences plus fiables que ceux issues suites aux imputations multiples et simples (Tableau 1). Ces résultats suggèrent que l'imputation par metaCART peut être plus efficace si la minimisation du biais est un critère prioritaire pour le méthodologiste.

Conclusion : La présente étude montre le potentiel d'une nouvelle approche d'imputation basée sur le principe du méta-apprentissage. Des procédures d'optimisation de performance (ex. Cross-Validation) peuvent être ajoutée à cette méthode afin de booster sa performance d'imputation. D'autres combinaisons de méta-algorithmes et imputeurs de base peuvent être explorées également. Notre travail n'est qu'une initiation à un travail important d'investigation de cette nouvelle approche proposée.

Mots clés : Imputation multiple , Imputation simple , Méta-apprentissage , Données manquantes , CART

Déclaration de liens d'intérêts : Les auteurs n'ont pas précisé leurs éventuels liens d'intérêts.

Tableau 1 : Bias d'estimation post-imputation (moyennes des différences des HRs obtenus sur 600 jeux de données complets et HRs obtenus sur les mêmes jeux de données après introduction des données manquantes et imputations)

Covariables*	knn	metacart	mice	micecart	micerf	missCforest	missForest
x1	0.05	0.03	-0.05	0.02	0.02	0.02	0.05
x2	0.07	-0.02	-0.11	-0.09	-0.03	0.08	0.07
x31	-0.04	0.04	-0.04	0.05	-0.09	0.20	0.11
x42	-0.05	0.06	-0.21	0.02	-0.07	0.17	0.10
x43	-0.14	0.05	-0.02	0.01	-0.07	0.18	0.11

* x1 est continue, x2 est continue et corrélée à x1, x3 et binaire et x4 et catégorielle à 3 modalités

https://doi.org/10.1016/j.respe.2023.101632

CO10.7 - Plateforme de données de vie réelle ODH: élaboration d'un observatoire du médicament en oncologie

N. Benhajkassen[1,*], L. Bosquet[1], M. Deniau[1], C. Bachot[2], T. Guesmia[1], V. Robert[1], V. Machuron[2], A. Martin[1]

[1] Unicancer, Direction des datas et partenariats, Le Kremlin-Bicêtre, France

[2] Roche SAS, Centre de données médicales et de médecine personnalisée, Boulogne-Billancourt, France
*Auteur correspondant
E-mail address: n-benhajkassen@unicancer.fr

Introduction : Les données cliniques et thérapeutiques « de vie réelle » ont pris une place importante dans l'évaluation des produits de santé et des stratégies thérapeutiques, imposant la création de nouvelles plateformes dynamiques qui s'appuient sur les méthodes de big data et de « machine learning ». Dans ce contexte et dans le cadre d'un partenariat public-privé ouvert sous gouvernance scientifique académique, le « Onco Data Hub » (ODH) est un entrepôt de données de Santé (EDS) géré par Unicancer et constitué de données extraites à partir des données immédiatement disponibles dans les établissements de santé (ETS) prenant en charge des patients atteints de cancer en France.

Méthodes : La plateforme centralise plus de 60 variables clés (caractéristiques patients et maladie, traitements reçus, biomarqueurs, suivi du statut vital…) nativement structurées et collectées automatiquement via les logiciels de préparation des traitements anticancéreux parentéraux (CHIMIO®…) mais également des données issues du dossier médical des patients nécessitant l'usage d'outils de traitement automatique (NLP, RegEx) pour extraire la donnée recherchée. L'objectif est d'obtenir une base unique, structurée et exhaustive adaptée à de nombreux usages : description des populations cibles, étude d'efficacité… La collecte des données s'organise autour de deux campagnes annuelles avec l'intégration des données de patients nouvellement pris en charge ainsi que la mise à jour des patients déjà inclus lors des collectes précédentes. Les patients ayant eu une administration d'une thérapie anticancéreuse injectable au cours des six mois précédant l'extraction sont sélectionnés à partir du logiciel de préparation des traitements anticancéreux. Pour la première phase du projet, nous avons restreint la sélection aux patients ayant un diagnostic de cancer broncho-pulmonaire ou du sein. Une généralisation aux autres cancers est prévue ultérieurement. Les données collectées font systématiquement l'objet d'un contrôle visant à valider la qualité et la cohérence des données.

Résultats : En 2022, 25 premiers ETS ont participé à l'acquisition des données. L'objectif est fixé à environ 60 ETS représentatifs (CHU, CH, CLCC, cliniques…) d'ici fin 2023. La base comptait, en novembre 2022, 30 953 patients traités dans l'un des 25 ETS contributeurs entre le 01/01/2020 et le 30/06/2022, dont 18 782 patients atteints d'un cancer du sein et 12 171 patients atteints d'un cancer broncho-pulmonaire. Les caractéristiques des patients, de leur maladie et de leur prise en charge sont déjà accessibles et seront présentées. Dans certains cas, la non-structuration des données a nécessité une collecte manuelle directement à partir du dossier patient source sur site. Un travail d'automatisation de la récupération de ces données est en cours avec les ETS permettant, en plus d'atteindre l'objectif d'automatisation complète du processus de recueil des données prévu pour ODH, de s'inscrire dans le processus d'optimisation des systèmes informatiques propre à chaque ETS.

Conclusion : Le programme ODH va contribuer à une meilleure compréhension des stratégies thérapeutiques « en vraie vie » en constituant à terme une nouvelle source de données récentes, riche, dynamique et représentative de la prise en charge des patients qui permettra de mener des projets de recherche académiques et de produire des données nécessaires aux laboratoires pharmaceutiques, aux institutions de santé impliquées dans l'évaluation des produits de santé et à la décision en santé publique.

Mots clés : Données de vie réelle; Cancer du sein; Cancer Broncho-pulmonaire; Epidémiologie; Base de données