Retour d'expérience concernant le data challenge épigénétique et médiation à large échelle d'Aussois

Auteurs par ordre alphabétique : Sophie Achard [1], Dylan Aïssi [2,3], Michael Blum [4], Kevin Caye [4], Florent Chuffart [2], Olivier François [4], Keurcien Luu [4], Florian Privé [4], Magali Richard [4].

[1] *Grenoble Image Parole Signal Automatique, CNRS UMR5216.*
[2] *Institut pour l'Avancée des Biosciences, Inserm U1209, CNRS UMR5309, Université Grenoble Alpes.*
[3] *CHU Grenoble Alpes.*
[4] *Université Grenoble Alpes, Laboratoire TIMC-IMAG, CNRS UMR5525.*

Résumé. La médiation est un outil statistique permettant d'identifier des liens de causalité entre des variables environnementales et un phénotype par l'inclusion d'une troisième variable : le médiateur. L'étude de la médiation est un réel enjeu statistique car les jeux de données étudiés présentent un grand nombre de médiateurs, souvent corrélés, et complexes de part leur nature biologique. Un *data challenge* a été organisé pour présenter l'analyse de médiation à des chercheurs de plusieurs disciplines, favoriser les échanges entre les participants et en évaluer les méthodes couramment utilisées. Cet article détaille les objectifs pratiques et pédagogiques de l'événement : la préparation des jeux de données, la conception d'un site web pour l'évaluation des épreuves, des sessions pratiques et pédagogiques, et une restitution du déroulement des épreuves. En conclusion, le *data challenge* d'Aussois a été une expérience transdisciplinaire enrichissante sur le plan scientifique et méthodologique.

Mots-clés. médiation, *data challenge*, épidémiologie, méthylation, science des données

Abstract. Mediation is a statistical method to identify biological pathways that relate environmental variables to health outcomes. The application of mediation is challenging due to the complexity of modern data sets: there is a large number of potential mediators to consider; potential mediators might be correlated; biological a priori should be incorporated. A *data challenge* has been organised to gather researchers from very different disciplines, to give introduction to mediation analysis, finally to increase interactions between participants to go beyond the state of the art through evaluation of different statistical methods. This paper describes the practical and methodological objectives of the *data challenge*: the design of the web interface for the evaluation; a mix of pedagogical courses and practical sessions; preparation of the datasets; and communication of the results. In conclusion, the *data challenge* has been a interdisciplinary event where participants acquired scientific knowledge about mediation analysis and got to know the data challenge format.

Keywords. mediation, *data challenge*, epidemiology, methylation, data science

1 Introduction

1.1 Question scientifique : médiation et épigénétique

L'analyse de médiation est une approche permettant d'identifier des voies biologiques qui interviennent dans la relation entre facteur environnemental et pathologie. Le principe de l'analyse de médiation est d'identifier, à partir d'une association entre un facteur environnemental et une pathologie (ou un endophénotype), une variable médiatrice associée à la fois au facteur environnemental et à la pathologie en prenant en compte l'effet du facteur environnemental.

Le modèle de médiation se prête bien à la prise en compte des effets épigénétiques sur les maladies [1]. En effet, certains facteurs environnementaux comme la nourriture, le tabac ou la pollution atmosphérique sont connus pour induire diverses pathologies et il a été proposé que ces facteurs environnementaux dérégulent l'expression génique par le biais de perturbation de facteurs épigénétiques [2]. L'épigénétique est un ensemble de mécanismes biologiques réversibles faisant intervenir des molécules pouvant se fixer sur le génome pour favoriser ou entraver la fixation des facteurs de transcriptions. Une des marques épigénétiques actuellement très étudiée à l'échelle épidémiologique est la méthylation de l'ADN. L'identification des voies biologiques par lesquelles interviennent les associations entre facteurs environnementaux et pathologies mettrait en évidence de nouvelles cibles thérapeutiques pour contrer les effets des facteurs environnementaux.

1.2 Objectifs du *data challenge*

Un *data challenge* autour de l'épigénétique et de la médiation à large échelle a été organisé les 7, 8 et 9 juin à Aussois [3]. Sa conception s'articule autour de trois principes : i) réunir un public hétérogène d'ingénieurs, de chercheurs et d'étudiants provenant de disciplines aussi diverses que l'informatique, la statistique, la physique, l'épidémiologie ou encore la biologie ; ii) plonger ces participants dans le contexte scientifique de la médiation en lui présentant les enjeux et les outils du domaine ; iii) créer une émulation scientifique entre les participants autour de deux jeux de données simulés afin d'évaluer la pertinence des méthodes.

2 Aspects pédagogiques du déroulement

L'articulation des séances pédagogiques et des épreuves pratiques est un élément clef d'un tel évènement. En effet, une bonne compréhension des enjeux et des méthodes du domaine favorise l'apprentissage et la participations aux épreuves proposées.

2.1 Sessions pédagogiques

Les sessions pédagogiques ont permis d'introduire les questions scientifiques, les méthodes et les outils utilisés lors des épreuves. Voici le détail de ces sessions, qui ont été réparties tout au long de l'événement :

- 1 séminaire scientifique de 1 heure : *Exposure misclassification in environmental epigenetics: is DNA-methylation a mediator or a biomarker?* par Linda Valeri (Harvard University).
- 3 mini-cours de 15 minutes : *Introduction to environmental health* par Rémy Slama (Institut pour l'Avancée des Biosciences, Grenoble), *Introduction to mediation analysis for environmental health* par Johanna Lepeule (Institut pour l'Avancée des Biosciences, Grenoble), *Omics data in biomedical research (SNP, methylation marks)* par Magali Richard (Université Grenoble Alpes).
- 2 présentations des épreuves de 30 minutes : *Statistical analysis and R packages for mediation analysis* par Michael Blum (Université Grenoble Alpes), *Accounting for confounding factors in association studies* par Olivier Francois (Université Grenoble Alpes).

Les tutoriels ont permis d'introduire la nature des données, quelques méthodes pour aborder les épreuves, ainsi que les logiciels recommandés pour implémenter les solutions (ici R [4]). Il avait été demandé au préalable aux participants de télécharger les logiciels et les jeux de données nécessaires via le dépôt GitHub de l'événement [5].

2.2 Sessions pratiques

Le format adopté pour le *data challenge* d'Aussois s'inspire fortement des compétitions en ligne organisées dans le contexte de la science des données (kaggle [6], Data Science Game [7], Challenge data [8]). La participation à ce genre d'épreuves est connue pour être très formatrice, si bien qu'il n'est pas rare de trouver la mention "une expérience kaggle est un plus" dans les offres d'emploi en data science. Et pour cause, ces compétitions permettent aux participants de développer des compétences techniques variées en analyse de données (traitement de données, programmation, développement de modèles de prédiction, interprétation, etc), et peuvent donc s'avérer très utiles pour toute personne travaillant dans le domaine de la bioinformatique ou de la biostatistique. L'avantage de

ces épreuves est qu'elles permettent rapidement aux participants de saisir les enjeux d'un problème avec lequel ils ne sont pas forcément familiers. Compte tenu des objectifs et de la durée de la formation, l'articulation de mini-épreuves autour des présentations orales s'est révélée être un choix pertinent.

L'espace web dédié à l'évaluation participe au bon déroulement du *data challenge*. Un site web a donc été mis en place afin de permettre aux participants d'évaluer leurs soumissions en ligne, de la même manière que sur les plateformes mentionnées ci-dessus. Par souci de simplicité (et par conviction personnelle), ce site a été entièrement réalisé avec l'outil R Shiny [9]. En pratique, pour l'utilisateur/trice, chaque soumission sur le site lui renvoie un score mesurant la puissance et le taux de fausses découvertes, lui permettant ainsi d'évaluer rapidement la pertinence de son modèle relativement à ces critères. Les participants sont organisés en équipe de 2 à 3 membres et soumettent leurs résultats sous l'identifiant de leur équipe. À chaque soumission, l'équipe est classée en fonction de son score, le classement est affiché sur un *leader board* accessible à tous.

3 Méthode : création des jeux de données

Durant les épreuves, les participants ont pour objectif de retrouver un sous ensemble de médiateurs qui interviennent dans la relation entre facteur d'exposition et phénotype. Les données ont été simulées suivant un modèle de médiation multiple [10].

Le premier jeu de données contient la valeur de 5 000 variables continues qui sont des médiateurs potentiels typés pour 500 individus. Parmi ces 5 000 variables, 500 variables sont influencées par un facteur d'exposition qui est lui même simulé. Un total de 60 variables est tiré au hasard parmi ces 500 variables pour constituer l'ensemble des médiateurs. Le trait de santé est ensuite défini par une combinaison linéaire des 60 médiateurs et d'un bruit gaussien. Comme il y a de la variabilité dans la simulation, la simulation a été répétée jusqu'à ce que les résultats des méthodes d'identification des médiateurs donne sur le jeu de données un résultat considéré comme typique.

Le second jeu de données s'appuyait sur des données réelles issues de l'étude de Vandiver et al. 2015 [11]. L'étude comportait 78 échantillons de tissus (derme et épiderme), un facteur d'exposition, l'âge et le genre des patients, et des données de méthylation Illumina 450k pour chaque échantillon. Pour des raisons pratiques liées à l'organisation des épreuves, un petit nombre de profils de méthylation normalisés a été extrait de la puce (taux de méthylation (valeurs bêta) de 1 496 sites du chromosome 1). Un phénotype binaire a ensuite été simulé pour chacun des 78 échantillons selon un modèle de risque binomial impliquant 19 médiateurs d'effet identique. Les médiateurs ont été choisis parmi 100 sondes dont les valeurs bêta avait été tout d'abord simulées, puis placées dans l'ensemble des profils de méthylation. Le modèle génératif prenait en compte l'association à la variable d'exposition, ainsi qu'un bruit corrélé de même matrice de covariance que les valeurs observées.

4 Participation et retour d'expérience

4.1 Statistiques de participation

Le *data challenge* a réuni 27 participants (francophones et non francophones), répartis dans les disciplines suivantes: statistique, épidémiologie, informatique, médecine et biologie. Lors des sessions pratiques, les participants ont constitué des équipes afin de répondre aux différentes épreuves (11 équipes). La transdisciplinarité de l'événement a favorisé la découverte d'approches inédites basées sur la combinaison de méthodes issues des différentes disciplines. Le principe de compétition interactive à l'aide de la mise à jour du *leader board* a significativement incité les équipes à soumettre leurs résultats de manière régulière et à activement chercher de nouvelles solutions plus performantes (en moyenne 30 soumissions par équipe). Chaque épreuve s'est soldée par une présentation du *leader board* final, accompagnée d'une restitution par chaque équipe des méthodes utilisées et des problèmes rencontrés.

4.2 Retour d'expérience

Les participants ont particulièrement apprécié les particularités suivantes de l'événement : i) l'aspect ludique apporté par le format compétitif, ii) l'intérêt indéniable de la pratique concrète dans l'apprentissage méthodologique, iii) les discussions scientifiques favorisées par la petite taille de la communauté, une session poster, le travail en équipe, et la présentation des résultats de chaque équipe lors des sessions de restitution, et iv) les approches originales apportées par une vision transdisciplinaire de la question scientifique abordée.

5 Discussion

Le *data challenge* d'Aussois a rempli ses objectifs qui étaient de réunir un public transdisciplinaire autour du contexte scientifique de la médiation et de créer une émulation scientifique afin de développer de nouvelles des méthodes. Les participants étaient satisfaits du format et des épreuves proposés. Nous souhaitons également mettre en avant dans cette discussion plusieurs points qui pourraient être améliorés lors de la réalisation de prochains évènements de ce type.

Nous avons observé que les participants pouvaient faire progresser leur score en apprenant directement de leurs soumissions précédentes et de la structure des jeux de données simulées. Il est important de prendre en compte cet aspect lors de création des jeux de données en amont de l'événement, et de bien maîtriser le modèle utilisé pour simuler les données. Pour éviter le surapprentissage, il nous semble également pertinent de préparer deux jeux de données pour chaque épreuve. Le premier jeu de données sert à l'apprentissage et permet de tester les méthodes développées et de voir son score s'afficher

sur le *leader board*. Le deuxième jeu de données sert à l'évaluation finale des différentes équipes. Ce processus est d'ailleurs généralement utilisé dans les évènements kaggle. Enfin, dans le but de rendre plus robuste l'interface web, il conviendrait d'utiliser une base de données pour l'enregistrement des utilisateurs ainsi que pour le stockage des soumissions, en lieu et place des Google sheets.

Le format *data challenge* ouvre des perspectives en terme d'enseignement des statistiques. Pour le bon fonctionnement de l'outil, il semble important de prendre en compte l'articulation des séances pédagogiques et des épreuves pratiques. En particulier, il faut veiller à ne pas présenter de contenu pédagogique une fois une épreuve lancée, les participants étant concentrés sur le développement de leur méthode. Du point de vue des intervenants, il peut être difficile de capter l'attention des participants, sollicités par différents médias (web, vidéo projecteur). En revanche, ces médias peuvent être avantageusement utilisés par l'intervenant. Cela peut prendre la forme des diapositives que les participants font défiler sur leurs ordinateurs personnels, ou des formes plus interactives offertes par l'usage d'un espace web adapté.

La diffusion des résultats et des avancées réalisées durant un tel évènement est une question difficile, étant donné le peu de temps et de disponibilité dont disposent les participants pour rédiger une communication scientifique. Nous travaillons actuellement sur plusieurs canaux de diffusion dont une publication dans le domaine de l'enseignement des statistiques, un post de blog relatant les faits marquants du séjour, et la rédaction d'un article d'opinion scientifique sur les méthodes développées et testées. Une autre manière de valoriser cette expérience est la diffusion d'un méta-package R contenant les jeux de données simulés ainsi que les restitutions de participants sous forme de vignettes.

Remerciements

Ce travail a été soutenu par le Labex PERSYVAL-Lab (ANR-11-LABX-0025-01) et par le Grenoble Alpes Data Institute qui est financé par l'Agence Nationale de la Recherche dans le cadre du programme "Investissements d'avenir" (ANR-15-IDEX-02).

Bibliographie

[1] "DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring", *Int J Epidemiol.* (2015 44(4): 1224-1237.
[2] Robert Feil & Mario F. Fraga, "Epigenetics and the environment: emerging patterns and implications", *Nature Reviews Genetics* (2012) 13, 97-109.
[3] Epigenetic & High-Dimension Mediation Data Challenge. June 7-9 2017, Aussois (73), France. https://data-institute.univ-grenoble-alpes.fr
[4] R cran https://cran.r-project.org
[5] Dépôt GitHub du *data challenge* https://github.com/bcm-uga/mediation-challenge

[6] kaggle https://www.kaggle.com
[7] Data Science Game https://www.datasciencegame.com
[8] Challenge data https://challengedata.ens.fr
[9] Shiny https://shiny.rstudio.com
[10] Zhang et al. "Estimating and testing high-dimensional mediation effects in epigenetic studies". *Bioinformatics* (2016) 32 (20):3150-3154.
[11] Vandiver et al. "Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin". *Genome Biology* (2015) 16:80.