HEC MONTRÉAL

Caractérisation de l'immunité hybride au SRAS-CoV-2 à partir de la Biobanque québécoise de la COVID-19

par
Jean-Frédéric Boulianne

Denis Larocque, Ph.D.
HEC Montréal
Codirecteur de recherche

Delphine Bosson-Rieutort, Ph.D.
Université de Montréal
Codirectrice de recherche

Sciences de la gestion
(Spécialisation Intelligence d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(M. Sc.)

Décembre 2024

Résumé

Depuis décembre 2019, le virus du SRAS-CoV-2 a causé, selon l'OMS, plus de 6,9 millions de décès et plus de 769 millions de contaminations (au 20 août 2023). Sachant que l'hôte n'est immunisé que sur une courte période et que différents vaccins ont été administrés à la population, il résulte de multiples combinaisons longitudinales d'infection, de réinfections et de vaccinations, générant l'immunité dite hybride. L'étude des différents profils de réinfection s'avère essentielle afin de générer des données probantes qui contribueront à guider l'optimisation des réponses de santé publique. L'objectif de l'étude est de regrouper les participants selon leur similarité d'infection, de réinfections et de vaccinations, puis de caractériser les groupes ainsi identifiés sur le plan de différents facteurs sociodémographiques et cliniques. Pour ce faire, des méthodes d'apprentissage automatique ont été appliquées sur les données de 318 patients adultes de la Biobanque québécoise de la COVID-19 recrutés entre avril 2020 et mars 2023, incluant des cas sévères et légers suivis sur une période maximale de trois ans. Il avait également comme sous-objectif d'explorer différentes combinaisons d'algorithmes d'apprentissage automatique et de mesures de dissimilarité afin de réaliser des regroupements en présence de « *patterns »* temporels complexes tout en recherchant un équilibre entre la performance algorithmique et l'interprétabilité. Les principaux résultats consistent en différents profils et leurs caractéristiques sociodémographiques et cliniques. Un algorithme qui semblait le plus pertinent pour les données et le contexte de l'étude a également été identifié. L'étude des caractéristiques et des séquences temporelles des groupes a notamment permis de mettre en lumière le rôle des politiques de vaccination dans les profils d'infection et de réinfection.

Mots clés : COVID-19, Coronavirus, Immunité hybride, caractérisation, BQC19

Méthodes de recherche : Apprentissage automatique, analyse de regroupement

Abstract

Since December 2019, the SARS-CoV-2 virus has caused, according to the WHO, more than 6.9 million deaths and over 769 million infections (August 20, 2023). Given that the host is only immunized for a short period and that different vaccines have been administered to the population, this results in multiple longitudinal combinations of infection, reinfection, and vaccination, generating so-called hybrid immunity. The study of different reinfection profiles is essential to generate evidence that will guide the optimization of public health responses. The aim of the study is to group participants according to their similarity of infection, reinfections, and vaccinations, and then to characterize the groups thus identified in terms of various sociodemographic and clinical factors. Machine learning methods were applied to data from 318 adult patients of the Biobanque Québécoise de la COVID-19 recruited between April 2020 and March 2023, including severe and mild cases followed up for up to three years. The project also had a sub-objective to identify machine learning algorithms and dissimilarity measures best suited for clustering in the presence of complex temporal patterns while seeking a balance between algorithmic performance and interpretability. The main results consisted of different profiles and their socio-demographic and clinical characteristics. One algorithm best suited to the data and the context of the study was also identified. In particular, the study of group characteristics and temporal sequences highlighted the role of vaccination policies in infection and reinfection profiles.

Keywords : COVID-19, Coronavirus, hybrid immunity, BQC19

Research methods : Machine learning, clustering

Liste des tableaux et des figures

Tableaux

Figures

Liste des abréviations

ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
BQC19	Biobanque québécoise de la COVID-19
CER	Comité d'éthique de la recherche
CHSLD	Centre d'hébergement de soins de longue durée
CIQ	Comité sur l'immunisation du Québec
DTW	*Dynamic time warping* [Déformation temporelle dynamique]
EPI	Équipement de protection individuelle
FRQS	Fonds de Recherche du Québec - Santé
HL7	*Health level 7*
INSPQ	Institut nationale de santé publique du Québec
MSSS	Ministère de la Santé et des Services sociaux
OM	*Optimal matching* [Méthodes d'appariement optimal]
OMS	Organisation mondiale de la Santé
PCR	*Polymerase chain reaction*
RI-RTF	Ressources intermédiaires et de type familial
RPA	Résidence privée pour aînés
SARS-CoV-2	*Severe acute respiratory syndrome coronavirus 2*

SCCAI	*Standard classification of countries and areas of interest*
SRAS-Cov-2	Coronavirus 2 du syndrome respiratoire aigu sévère
TdeS	Travailleur de la santé

Remerciements

J’aimerais d’abord remercier mes trois superviseurs, pour leur soutien, leur confiance et leurs précieux conseils. Plus précisément, je souhaite formuler ma reconnaissance au professeur Simon Rousseau de m’avoir offert l’occasion de participer à ce projet sur l’immunité hybride et de collaborer avec la Biobanque québécoise de la COVID-19.

Au professeur Denis Larocque, je souhaite adresser mes sincères remerciements d’avoir accepté d’agir à titre de codirecteur, de s’être joint au projet et d’y avoir apporté son expertise.

À la professeure Delphine Bosson-Rieutort, ma codirectrice, qui m’a accueilli au sein de son équipe de recherche malgré mon appartenance à une tierce université, je tiens à lui témoigner ma plus profonde gratitude pour son soutien, sa rigueur, sa passion, sa disponibilité, sa patience, mais également pour l’initiation au domaine de la recherche et pour les nombreuses occasions d’apprentissage connexes au projet.

Je tiens également à remercier mes collègues de l’équipe de recherche EDoS pour leur soutien, les échanges stimulants, mais également pour la création d’un espace d’apprentissage bienveillant. Alexandra et Juliette, un immense merci pour la complicité et pour votre capacité à insuffler la bonne humeur.

À Simon C., un tendre merci pour ton inéluctable soutien, mais également pour ton rôle d’intermédiaire inopiné qui, par un hasard d'un des chemins de ta vie, m’a mis en contact avec Delphine. Merci de croire en moi.

À mes proches et amies, un remerciement spécial pour votre appui, mais également pour la compréhension des efforts et des renonciations que la réalisation de cette aventure tout en travaillant à temps plein a nécessité.

En guise de conclusion, j'aimerais exprimer une pensée particulière pour Josée qui, à son insu et malgré le temps et la distance, joue un rôle de premier plan dans mon parcours. Je termine ainsi avec une citation d'un de tes auteurs favoris qui, de mon point de vue, est plus qu'à propos lorsque l'on songe au contexte pandémique du projet, mais également à notre réseau de la santé et des services sociaux et à ses nombreuses transformations passées… et à venir…

« La résilience, c'est l'art de naviguer dans les torrents » - Boris Cyrulnik

Ce travail a été rendu possible grâce au partage de données et d'échantillons de la Biobanque québécoise de la COVID-19, financée par le Fonds de recherche du Québec - Santé, Génome Québec, l'Agence de la santé publique du Canada et, depuis mars 2022 le ministère de la Santé et des Services sociaux. Nous remercions tous les participants à la BQC19 pour leur précieuse contribution.

https://www.quebeccovidbiobank.ca

Chapitre 1 Introduction

1.1 Contexte

En décembre 2019, des cas de pneumonies de cause non identifiée sont répertoriés en Chine et un nouveau coronavirus, lié au marché Huanan dans la ville de Wuhan, est alors identifié (Crits-Christoph *et al.*, 2024; Gostin Lawrence et Gronvall Gigi, 2023; Huang *et al.*, 2020; Worobey *et al.*, 2022; Zhu *et al.*, 2020). Dès janvier 2020, l'Organisation mondiale de la Santé (OMS) confirme une transmission interhumaine pour ce coronavirus dénommé SRAS-CoV-2 (World Health Organization, 2020a). Cette même organisation en déclarera l'aspect pandémique global le 11 mars 2020, soit moins de trois mois plus tard (Arabi *et al.*, 2023; Institut national de santé publique du Québec, s.d.; Tremblay *et al.*, 2021). En moins de vingt ans, il s'agit, après le SRAS-CoV-1 en 2002 en Chine et le MERS-CoV en 2012 dans la péninsule arabique, de la troisième menace sanitaire mondiale liée à un coronavirus (Bonny *et al.*, 2020; Gralinski et Menachery, 2020; Hu *et al.*, 2021). Cette dernière déclenchera toutefois une période pandémique socialement marquante à l'échelle planétaire entraînant la pandémie la plus importante depuis la grippe espagnole de 1918 (Lapuente, Winkler et Tenbusch, 2024) avec, en date d'avril 2024, plus de 7 millions de décès et plus de 770 millions de contaminations (Livieratos, Gogos et Akinosoglou, 2024; World Health Organization, 2024).

La pandémie a démontré que les individus ayant contracté la maladie ne sont pas immunisés de façon permanente et peuvent être infectés de nouveau après une courte période. En effet, ils sont susceptibles d'être infectés de nouveau après une période relativement courte d'environ 7 à 12 mois (Misra et Theel, 2022; Rodriguez Velásquez *et al.*, 2024). Cette variabilité dans le temps de réinfection peut s'expliquer par les sensibilités individuelles, la sévérité de l'infection (He *et al.*, 2021; Röltgen *et al.*, 2020), mais également par le fait que le virus SRAS-CoV-2 a évolué à travers le temps, impliquant la circulation de différents variants au sein des populations. Ces diverses mutations génèrent des réponses immunitaires différentes qui impactent l'immunité des individus ayant contracté la maladie.

Au Québec, en réponse à cette menace, le Fonds de recherche du Québec – Santé (FRQS), en collaboration avec Génome Québec, a constitué la Biobanque québécoise de la COVID-19 (BQC19) (McGill University, 2020; Scientifique en chef du Québec, 2020; Tremblay *et al.*, 2021). Initiée le 1er avril 2020, elle a pour mission de « s'assurer que les scientifiques ont accès au matériel biologique et aux données nécessaires à leurs efforts de recherche sur la COVID-19 » (Groupe de travail sur l'immunité face à la COVID-19, 2021; Tremblay *et al.*, 2021). Ainsi, issue d'un réseau d'universités et d'hôpitaux du Québec, la BQC19 a colligé, durant toutes les vagues d'infection de la pandémie de SRAS-CoV-2, un volume important de données cliniques et biologiques ainsi que certaines données clinico-administratives pour de nombreux participants, incluant les différentes manifestations cliniques de la maladie. Ces données longitudinales constituent une source remarquable d'information pour en apprendre davantage sur la réponse immunitaire des participants.

1.2 Objectifs

Dans ce contexte, nous souhaitons identifier et caractériser les différents profils de réinfection au SRAS-CoV-2 de manière à étudier l'immunité hybride, c'est-à-dire l'immunité octroyée par une combinaison d'infection, de réinfection et de vaccination.

Pour ce faire, nous proposons d'exploiter les données de la BQC19 à l'aide de techniques d'apprentissage automatique afin d'étudier l'immunité hybride au SRAS-COV-2 et, par le fait même, déterminer comment l'identification de profils de réinfection peut supporter la prise de décision en termes de gestion et de politiques de santé afférentes. Plus spécifiquement, il était question de :

1. Regrouper les individus selon leur similarité (« *pattern* ») de vaccination, d'infection et de réinfection;
2. Caractériser ces groupes en termes de facteurs sociodémographiques et cliniques.

Le projet avait également comme sous-objectifs :

1. La préparation, le nettoyage et la transformation des données de la BQC19 de manière à les optimiser pour les analyses de données temporelles;
2. L'exploration d'algorithmes d'apprentissage automatique et de mesures de dissimilarité pour réaliser des regroupements en présence de « patterns » temporels complexes, en recherchant un équilibre entre performance algorithmique et interprétabilité.

Le mémoire est structuré en sept chapitres distincts. Ce premier chapitre constitue l'introduction et présente les objectifs de la recherche. Le deuxième chapitre offre une revue de la littérature, abordant le contexte de la COVID-19 ainsi que les concepts associés de vaccination, d'infection, de réinfection et d'immunité hybride. Elle présente également l'apprentissage automatique, plus spécifiquement les analyses de regroupement. Le troisième chapitre présente la méthodologie adoptée pour le projet de recherche. Étant un mémoire par article, le quatrième chapitre est consacré à l'article préparé pour la *Revue canadienne de santé publique*, lequel présente les résultats de l'algorithme et de la matrice de dissimilarité ayant le mieux performé sur les données de la BQC19. Le cinquième chapitre présente à titre de résultats supplémentaires, les algorithmes explorés, mais qui ne sont pas inclus dans l'article. Le sixième chapitre propose une discussion approfondie et une interprétation des résultats obtenus, suivi d'une conclusion au septième chapitre. Enfin, le mémoire se conclut par une liste des références consultées.

Chapitre 2 Revue de la littérature

2.1 COVID-19

Acronyme de l'anglais de « *coronavirus disease-2019* » (Karia *et al.*, 2020), la COVID-19 est le nom officiel donné par l'OMS pour désigner la maladie causée par le virus SRAS-CoV-2 (World Health Organization, 2020b). Initialement nommé nCoV-2019 avant sa nomination par le comité international de taxonomie des virus (Bonny *et al.*, 2020), le « *Severe Acute Respiratory Syndrome Coronavirus-2* » (SARS-CoV-2 en anglais) est un virus de la famille des coronavirus (Livieratos *et al.*, 2024; Misra *et al.*, 2022; World Health Organization, 2020b). Cette dernière peut infecter différents animaux et cause, chez l'être humain, des infections respiratoires de gravité modérée à sévère (Hu *et al.*, 2021; National Institute of Allergy And Infectious Diseases, s.d.; World Health Organization, s.d.-a).

Répertorié en décembre 2019 dans la ville de Wuhan en Chine (Gostin Lawrence *et al.*, 2023; Huang *et al.*, 2020; Worobey *et al.*, 2022; Zhu *et al.*, 2020), le SRAS-CoV-2 se propagea rapidement à travers la planète. Le 3 janvier 2020, l'OMS rapportait un nombre de 44 cas en Chine, lequel passa, le 27 janvier, à 2 798 cas, dont 37 en provenance de 11 pays autres que la Chine (World Health Organization, 2020c, d). Le 1er février 2020, ce nombre se chiffrait à près de 12 000 cas et 23 pays étaient désormais touchés (132 cas) (World Health Organization, 2020e). Quinze jours plus tard, le nombre de contamination atteint plus de 50 000 cas dont 526 à l'extérieur de la Chine (25 pays) pour un peu plus de 1 500 décès (World Health Organization, 2020f). C'est ainsi que moins de trois mois après les premières contaminations humaines, le caractère alarmant du niveau de propagation amena l'OMS à en déclarer la pandémie mondiale le 11 mars 2020 (World Health Organization, 2020a). Le nombre de cas se chiffrait alors à plus de 118 000 dont un peu moins de 37 500 externes à la Chine répartis dans 113 pays pour un total d'un peu plus de 4 200 décès globalement (World Health Organization, 2020g).

Bien que la majorité des symptômes s'apparentaient à ceux d'une infection respiratoire typique, incluant de la fièvre, de la fatigue et de la toux (Hu *et al.*, 2021; Zhu *et al.*, 2020),

une proportion significative d'infection progressait vers une forme plus sévère, voire critique, de la maladie, pouvant impliquer de la dyspnée, un syndrome de détresse respiratoire aiguë et la défaillance de multiples organes (Gralinski *et al.*, 2020; Hu *et al.*, 2021; Lapuente *et al.*, 2024). En surplus de la forme aiguë de la maladie, l'infection pouvait mener à des problèmes de santé persistants (p. ex. fatigue, essoufflement, problèmes cognitifs) désigné syndrome post-COVID-19 aussi appelé COVID longue (Gouvernement du Canada, 2024; Lapuente *et al.*, 2024; World Health Organization et Kryuchkov, 2022).

La célérité avec laquelle la maladie s'est propagée et ses répercussions sur les hospitalisations ont notamment mis une pression sur les systèmes de santé, tant du point de vue de ses ressources que de son offre de services. La pandémie a notamment affecté les chaînes de production et de distribution mondiales impactant ainsi l'approvisionnement de divers produits, dont les équipements de protection individuelle (EPI) servant à protéger les travailleurs de la santé. Les organisations des systèmes de santé des pays industrialisés recherchaient les mêmes produits en réponses aux besoins croissants en EPI dans un contexte de sources d'approvisionnement limitées et incertaines (Beaulieu *et al.*, 2021). Cette incertitude a accentué l'état de stress du personnel de soins déjà sous tension par le contexte pandémique (Beaulieu *et al.*, 2021). Ces derniers ont également vu leurs conditions d'exercice changer (cohortage, réorientation vers les secteurs critiques, etc.) et leurs nombres d'heures de travail augmenter. L'Institut canadien d'information sur la santé rapporte, pour la période 2020--2021, que 236 000 travailleurs de la santé, soit une proportion de 21%, ont réalisé des heures supplémentaires avec par semaine en moyenne 8,2 heures rémunérées et 5,8 heures non rémunérées (Canadian Institute for Health Information, 2021b).

Du côté de l'offre de services, les systèmes de santé ont dû s'adapter et trouver une balance entre la prise en charge des patients atteints de la COVID-19 et les patients avec d'autres problèmes de santé (Canadian Institute for Health Information, 2021a). Cette balance ne s'est toutefois pas faite sans impact sachant qu'un recul dans les services de santé offerts par les médecins du Canada est constaté en 2020-2021 à la hauteur de 7,1 % pour la médecine familiale et de 8,9 % pour la médecine spécialisée. Les paiements aux

médecins ont d'ailleurs diminué pour la première fois en 20 ans (2%) (Canadian Institute for Health Information, 2021b). Les traitements vitaux et urgents ont été priorisés et bon nombre de chirurgies ont été reportées, soit une différence de 560 000 chirurgies de moins pendant les 16 premiers mois de la pandémie par rapport à l'année précédente (Canadian Institute for Health Information, 2021a).

Les systèmes de santé ne sont pas les seuls à s'être adaptés et avoir évolué pendant la pandémie. En effet, pour se propager dans la population, le virus se réplique dans les cellules des hôtes infectés et il arrive que des erreurs de copie de son matériel génétique, les mutations, surviennent (Institut national de santé publique du Québec, 2023). Ce processus est normal pour les coronavirus et génère, lorsque que plusieurs virus comportent la mutation, la création de nouvelles lignées de virus appelés variants. Les virus produisent donc souvent des variants, mais ils ne deviennent préoccupants en matière de santé publique que lorsque la mutation affecte de façon importante la transmissibilité, la virulence, l'efficacité des vaccins ou les tests diagnostiques (Agence ontarienne de protection et de promotion de la santé, 2023). Pendant la pandémie, cinq variants principaux ont été considérés comme préoccupants par l'OMS, à savoir Alpha (lignée[1] B.1.1.7), Beta (lignée B.1.351), Delta (lignée B.1.617.2), tous trois détectés en décembre 2020 respectivement au Royaume-Uni, en Afrique du Sud et en Inde, suivi de Gamma (lignée P.1) rapporté pour une première fois au Brésil en janvier 2021 et finalement de Omicron (lignée B.1.1.529) identifié en novembre 2021 en Afrique du Sud (Cascella *et al.*, 2024).

L'apparition de ces variants coïncide également avec la recrudescence d'infection dans la population. Bien qu'ils expliquent en grande partie l'apparition de ces vagues (Dutta, 2022), c'est-à-dire des cycles d'augmentation rapide et soutenue de cas suivi d'une diminution, ils n'en sont pas les seuls facteurs responsables. Les politiques de santé publique, le comportement de la population, la durée de l'immunité et la période de l'année sont également des exemples d'éléments qui interviennent dans la dynamique des

[1] Les lignées réfèrent à la nomenclature Pango qui permet de standardiser l'appellation des lignées du SRAS-CoV-2 pour une utilisation par les chercheurs et les organisations de santé publique à travers le monde pour la recherche et le suivi de la transmission et de la propagation du virus (Rambaut et al., 2020).

vagues épidémiques (Maragakis, 2021). Conséquemment, la temporalité des vagues varie entre les territoires en fonction de ces multiples facteurs. Il n'y a donc pas de périodes universelles pour marquer le début et la fin des vagues de la pandémie. Au Québec, l'Institut national de santé publique du Québec (INSPQ) a établi qu'il y en avait eu sept[2] durant la pandémie de COVID-19 (Institut national de santé publique du Québec, 2024).

En somme, la COVID-19 constitue un enjeu de santé publique important considérant l'absence d'immunité préexistante au SRAS-CoV-2 chez les êtres humains (Misra *et al.*, 2022), la vitesse de propagation du virus, les risques afférents pour la santé et l'immunité susceptible d'être altérée par les multiples mutations. En date d'avril 2024, elle est responsable de plus de 7 millions de décès et plus de 770 millions de contaminations à travers le monde (Livieratos *et al.*, 2024; World Health Organization, 2024). Cette pandémie a imposé la mise en œuvre rapide de solutions pour limiter la propagation du virus, telle que des mesures d'isolement, mais également des efforts mondiaux pour développer rapidement un vaccin afin de protéger les individus et de préserver la capacité des systèmes de santé à soigner la population.

2.2 Vaccination COVID-19

La vaccination figure parmi les découvertes les plus importantes de la médecine (Canoui et Launay, 2019). Existant depuis la fin du 18e siècle (Guimaraes *et al.*, 2015; Plotkin, 2014), la vaccination réduit significativement la morbidité et la mortalité associées à plusieurs maladies (Canoui *et al.*, 2019; Guimaraes *et al.*, 2015; World Health Organization, s.d.-c). Selon l'Organisation mondiale de la Santé, elle évite chaque année 3,5 à 5 millions de décès découlant de plus de 20 maladies infectieuses majeures comme la diphtérie, le tétanos et la coqueluche, pour ne nommer que celles-ci (World Health Organization, s.d.-c). Elle a même fortement contribué à l'éradication de la variole (Guimaraes *et al.*, 2015; Henderson, 2011; Strassburg, 1982). Conséquemment, elle

[2] Vague 1 du 25 février au 11 juillet 2020; Vague 2 du 23 août 2020 au 20 mars 2021; Vague 3 du 21 mars au 17 juillet 2021; Vague 4 du 18 juillet au 4 décembre 2021; Vague 5 du 5 décembre 2021 au 12 mars 2022; Vague 6 du 13 mars au 28 mai 2022; Vague 7 à partir du 29 mai.

favorise le contrôle des maladies infectieuses dans la plus grande partie de monde (Canoui *et al.*, 2019).

Le principe de la vaccination est d'introduire dans un organisme une préparation antigénique (vaccins) afin d'induire une protection contre une bactérie ou un virus responsable d'une maladie infectieuse visant à éviter la survenue de la maladie ou d'en atténuer les manifestations cliniques (Canoui *et al.*, 2019; Ministère de la Santé et des Services sociaux, 2020). L'objectif est ainsi de générer la production d'anticorps chez l'hôte, comme s'il était exposé à la maladie, afin de générer une forme d'immunité en prévision d'une rencontre ultérieure avec l'agent infectieux (Gouvernement du Québec, 2023; Ministère de la Santé et des Services sociaux, 2020; World Health Organization, s.d.-b). Selon l'agent pathogène et le vaccin, les effets sur la réponse à cette rencontre ultérieure peuvent être différents allant de la réduction de la gravité de la maladie à une immunité complète permanente en passant par une immunité temporaire nécessitant des doses de rappel.

En matière de COVID-19 et, particulièrement dans le contexte pandémique associé, la vaccination poursuit ce même objectif de protéger la population. Les efforts mondiaux de même que les technologies plus récentes (vaccins composés d'acide nucléique (ADN, ARN), vecteur viral et protéine recombinante) ont favorisé le développement rapide de vaccins pour la COVID-19 (Koirala *et al.*, 2020). Bien qu'il ait été établi que les vaccins contre la COVID-19 ne préviennent pas entièrement l'infection, ils en réduisent considérablement la sévérité et en diminuent ainsi les manifestations sévères menant aux hospitalisations et aux décès (Niklas Bobrovitz *et al.*, 2023; Diani *et al.*, 2022; Livieratos *et al.*, 2024; Misra *et al.*, 2022). Moghadas *et al.* (2021), dans leur étude de l'impact de la vaccination sur la pandémie aux États-Unis, ont démontré que la vaccination a permis de diminuer de 65,5 % les hospitalisations aux soins intensifs et de 63,5 % les autres hospitalisations. Dans la seule première année de vaccination de la pandémie, le nombre de décès évités par les vaccins est estimé près de 20 millions (Arabi *et al.*, 2023). La vaccination a donc joué un rôle crucial dans le contexte pandémique, autant en termes de protection de la population que de diminution de la pression sur les systèmes de santé.

La vaccination a notamment permis de protéger les populations vulnérables des effets graves de la maladie. Les avis du Comité sur l'immunisation du Québec (CIQ) démontrent que les stratégies de vaccination ont évolué au fil de la pandémie selon différents facteurs, dont la disponibilité des vaccins, mais sans pour autant diminuer l'importance accordée à ces populations (Comité sur l'immunisation du Québec, 2020, 2021; Comité sur l'immunisation du Québec *et al.*, 2024; Comité sur l'immunisation du Québec *et al.*, 2022). On compte, parmi ces populations, les personnes résidant dans les centres d'hébergement de soins de longues durées (CHSLD), les résidences privées pour aînés (RPA) ou les milieux de vie collectifs similaires, les personnes âgées de 60 ans et plus, les personnes âgées de plus de 6 mois immunodéprimées, dialysées ou avec une maladie chronique, les personnes enceintes, les adultes vivant en région éloignée et isolée et, finalement, les travailleurs de la santé (TdeS) (Aleem, Akbar Samad et Vaqar, 2024; Gouvernement du Québec, 2024).

Au Québec, les premiers vaccins ont été administrés dans la semaine du 13 décembre 2020. Au total, 4 843 doses ont été inoculées, dont 92,7 % à des travailleurs de la santé et 5,1 % à des résidents des CHSLD. En janvier 2021, la vaccination massive de ces deux groupes prioritaires a continué (215 837 doses, dont 64,5 % pour les TdeS et 13,4 % pour les résidents des CHSLD) suivi par les résidents des RPA en février (202 078 doses, dont 22,3 % pour les TdeS et 51,4 % pour les résidents en RPA). Deux ans après l'annonce par l'OMS de la désignation de pandémie mondiale, soit le 11 mars 2022, près de 19 millions de doses avaient été administrés à la population, incluant les différentes doses de rappel recommandées par les autorités.

2.3 Infection, réinfection et immunité hybride

Une infection désigne l'invasion d'un organisme par un agent pathogène qui y prolifère et affecte ses tissus (Bush, 2022; National Cancer Institute, s.d.; National Institutes of Health (US), 2007; National Library of Medecine, s.d.). L'agent infectieux peut alors être transmis à d'autres organismes via différents modes de transmission. Face à cette invasion, le corps de l'hôte génère une réponse de défense contre l'agent infectieux (National Institutes of Health (US), 2007). Ce mécanisme lui confèrera, pour une période temporelle variable, une immunité acquise dite naturelle lui permettant, en cas

d'exposition ultérieure au pathogène, une réponse immunitaire plus rapide (Ministère de la Santé et des Services sociaux, 2018) l'empêchant de développer la maladie ou d'en atténuer la gravité.

L'hôte précédemment infecté par un virus (infection primaire) puis guéri peut parfois être de nouveau réinfecté par le virus (infection secondaire) (Centers for Disease Control and Prevention, 2024). Il est alors question de réinfection. Dans le contexte de la COVID-19, il n'y a pas, à l'heure actuelle, de consensus quant au délai minimum entre les infections primaire et secondaire pour considérer être en présence d'une réinfection (Chen *et al.*, 2024). Plusieurs publications considèrent que, d'un point de vue épidémiologique, une réinfection se définit par un test PCR positif survenu plus de 90 jours suivant le test positif de l'infection primaire (Centers for Disease Control and Prevention, 2024; Pilz *et al.*, 2022; Yahav *et al.*, 2021). Dans leur revue systématique et méta-analyse, Chen *et al.* (2024) soulèvent que d'autres études utilisent plutôt un intervalle de 30 jours pour considérer une réinfection. Dans une étude longitudinale sur la population du Qatar, Chemaitelly *et al.* (2024) proposait plutôt un intervalle de 40 jours comme alternative au plus conventionnel 90 jours après avoir exposé que presque tous les tests positifs dans les 15 jours de l'infection primaire étaient attribuables à l'infection initiale, alors que presque tous les tests positifs dépassant les 30 jours étaient plutôt attribuables à une réinfection. Bien que dès 2020, le Centre européen de prévention et de contrôle des maladies appelait à une définition commune (European Centre for Disease Prevention and Control, 2020), ces éléments démontrent l'absence de délai universel pour définir la réinfection.

Finalement, l'immunité est dite hybride lorsque l'immunisation est conférée par une combinaison d'une infection et de vaccination (Boaventura, Cerqueira-Silva et Barral-Netto, 2023; N. Bobrovitz *et al.*, 2023; Jacobsen *et al.*, 2023; Livieratos *et al.*, 2024; Misra *et al.*, 2022; Pilz *et al.*, 2022; Rodriguez Velásquez *et al.*, 2024; Stulpin, 2023). Cette dernière revêtait un intérêt particulier afin de comprendre comment les différentes expériences d'infection et de vaccination pouvait influencer la réponse immunitaire de la population. Ainsi, dans ce contexte où la séquence, la durée et la temporalité des événements pouvait avoir une importance majeure, le recours à des approches issues de la science des données, tels que les méthodes d'apprentissage automatique, s'avérait une

perspective intéressante pour analyser des données multidimensionnelles et temporelles complexes pour lesquelles les relations sont non linéaires et difficile à étudier via des méthodologies classiques.

2.4 Analyses de regroupement (*clustering*)

L'apprentissage automatique est un champ des sciences de l'informatique qui étudie les algorithmes et les techniques informatisées visant à résoudre des problèmes complexes de manière automatisée, c'est-à-dire sans que la tâche ait été programmée explicitement via les méthodes de programmation conventionnelle (Mahesh, 2020; Rebala, Ravi et Churiwala, 2019). Les méthodes d'apprentissage automatique incluent des algorithmes supervisés et non supervisés. Les premiers utilisent des données dont le résultat est déjà connu pour établir des règles de décision et classifier de nouvelles observations alors que les seconds utilisent des données sans à priori afin de découvrir des structures ou schémas sous-jacents. Les analyses de regroupement, *clustering* en anglais, sont une famille de ces méthodes d'apprentissage automatique non supervisées dont l'objectif est d'identifier des groupes homogènes dans les observations basées sur leur similitude ou la distance entre elles par rapport aux observations des autres groupes (Hastie *et al.*, 2009; Rodriguez *et al.*, 2019). Ces méthodes visent donc à définir des classes à partir des données sans a priori (Bhatia et Chiu, 2017; Hirano, Sun et Tsumoto, 2004; Nayyar, Gadhavi et Zaman, 2021; Rodriguez *et al.*, 2019). Parmi les méthodes les plus communément utilisées figurent les classifications hiérarchiques, les partitionnements en k-moyennes, les regroupements basés sur des modèles et les regroupements basés sur la densité (Bhatia *et al.*, 2017). Comme le mentionnent Nayyar *et al.* (2021), les services de santé peuvent être améliorés drastiquement en exploitant le potentiel de l'apprentissage automatique. Pour le domaine de la santé, une des applications les plus répandues est de mettre en lumière des groupes d'observations qui peuvent mener à la découverte de nouvelles maladies ou de nouveaux éléments (Hirano *et al.*, 2004), supportant ainsi la prise de décision fondée sur les données en termes de diagnostics et de traitements (Nayyar *et al.*, 2021). D'un point de vue de santé publique, l'apprentissage automatique est de plus en plus utilisé et ses applications sont nombreuses. On compte parmi les exemples la prédiction de retombées au niveau

populationnel des politiques de santé, de l'utilisation des services de santé et de l'incidence de certaines maladies (Alanazi, 2022).

Bien que l'apprentissage automatique offre des perspectives prometteuses dans le domaine de la santé et apporte parfois des solutions intéressantes, il présente également des limites et des défis. Ozaydin, Berner et Cimino (2021), dans leur étude sur son utilisation appropriée au domaine de la santé, relevaient plusieurs limites inhérentes à cette approche, notamment la présence de biais, la difficulté d'interprétation et d'explicabilité de certains algorithmes, ainsi que la problématique de la maintenabilité et de la validité des modèles dans le temps. À titre d'illustration, Wynants et al. (2020) ont mis en évidence, dans leur revue systématique portant notamment sur 606 modèles pronostiques liés à la COVID-19, que 593 d'entre eux (97,8 %) présentaient un risque élevé de biais, principalement en raison de la taille limitée des échantillons et de l'absence de validation externe des modèles. Cette problématique ne se restreint pas aux modèles prédictifs et pronostiques, mais s'étend également aux analyses de regroupement tel que présenté dans un commentaire de Van Smeden, Harrell et Dahly (2018) sur les travaux de Ahlqvist et al. (2018).

Malgré ces limites, l'apprentissage automatique présente un potentiel intéressant pour le domaine de la santé. Il convient ainsi de prendre en compte ces différents défis lors de l'utilisation de l'apprentissage automatique afin d'en réduire les risques, voire de les éliminer, pour tirer profit pleinement des bénéfices qu'il peut apporter. Parmi les solutions visant à atténuer les impacts de ces limitations figure d'assurer une utilisation compatible entre la problématique, la réponse à y apporter et l'apprentissage automatique. Entre autres, il est requis de veiller à la taille des échantillons ainsi qu'à leur représentativité, mais également à l'équilibre entre la performance des modèles et leur interprétabilité (Lee et Shin, 2020). Dans cette perspective, Ozaydin et al. (2021) soulignent l'importance d'« ouvrir la boîte noire », mettant ainsi de l'avant la nécessité d'une explicabilité suffisante des modèles. De surcroît, le choix d'un algorithme adapté à la problématique constitue une importante mesure d'atténuation face à ces défis. Ce choix doit être guidé par plusieurs facteurs, incluant le volume, la nature, la diversité et la vélocité d'évolution des données (Lee et al., 2020). En effet, le recours à l'apprentissage automatique s'avère

particulièrement pertinent par rapport aux méthodes classiques (par exemple les régressions) face aux données complexes et temporelles où les relations sont potentiellement non linéaires, dans un contexte de grand volume de données pour peu d'individus ou, encore, en présence de multicolinéarité, de données hétérogènes et multidimensionnelles. Finalement, les projets impliquant l'apprentissage automatique devraient prévoir la validation des résultats dans leur méthodologie. À défaut, les résultats pourraient plutôt être considérés comme préliminaires et les conclusions nuancées en conséquence pour éviter de potentielles surinterprétations. Walsh et al. (2021) proposaient à cet effet un ensemble de recommandations intitulés DOME dont l'adoption généralisée contribuerait, selon leurs travaux, à améliorer l'évaluation et la reproductibilité de l'apprentissage automatique. Ces recommandations portent sur quatre sphères des projets d'apprentissage automatique à savoir les **d**onnées, l'**o**ptimisation, les **m**odèles et l'**é**valuation.

En conclusion, l'apprentissage automatique présente un potentiel intéressant pour le domaine de la santé, bien que son utilisation ne soit pas exempte de défis. Une adéquation avec la problématique s'avère requise de même qu'un choix éclairé en matière d'algorithmes, lequel devrait entre autres être appuyé sur les données disponibles. Ce choix doit également tenir compte d'un équilibre entre la performance et l'interprétabilité des modèles. Les algorithmes du présent projet s'inscrivent dans cette volonté d'adéquation avec les données et d'interprétabilité, aussi bien des algorithmes que de leur extrant.

Chapitre 3 Matériel et méthodes

Afin de répondre à notre problématique qui était d'identifier et de caractériser les différents profils de réinfection au SRAS-CoV-2 de manière à étudier l'immunité hybride, la Biobanque québécoise de la COVID-19 a servi de source de données sur laquelle nous avons appliqué des méthodes d'apprentissage automatique après un nettoyage approfondi des données. Ce chapitre explicite le détail de ces différentes étapes de préparation et d'analyses.

3.1 Biobanque québécoise de la COVID-19

La Biobanque québécoise de la COVID-19 est une banque de données multicentrique composée d'un réseau de 11 hôpitaux du Québec et de 5 établissements académiques partenaires (Tremblay *et al.*, 2021). Cette initiative panprovinciale a été annoncée le 26 mars 2020 par le FRQS et Génome Québec, avec le soutien financier de l'Agence de la santé publique du Canada pour une entrée en opération le 1er avril de la même année (Tremblay *et al.*, 2021). Depuis 2022, le Ministère de la Santé et des Services sociaux (MSSS) s'est également joint à titre de bailleur de fonds (Biobanque québécoise de la COVID-19, 2022b). En surplus de sa mission de rendre disponibles des données sur la COVID-19 pour les chercheurs du domaine clinique, elle poursuit également l'objectif d'améliorer les efforts de recherche dans les domaines de la prévention, l'épidémiologie et la gestion populationnelle de la maladie (Tremblay *et al.*, 2021).

La BQC19 contient deux cohortes, l'une concernant les patients ayant développé une maladie sévère (cohorte hospitalisée) et l'autre composée de patients ayant développé une forme peu sévère de la maladie (cohorte ambulatoire). Plusieurs jeux de données permettent de documenter les caractéristiques des participants, les différents événements de leur trajectoire (cohorte hospitalisée) et certaines données biologiques (ARN, ADN, sérum, plasma et cellules mononucléées du sang périphérique, etc.) (Biobanque québécoise de la COVID-19, 2022a). Ils incluent également des suivis longitudinaux d'une durée de 24 mois suivant l'hospitalisation (cohorte hospitalisée) ou le dépistage par test PCR (cohorte ambulatoire).

3.1.1 Population étudiée et période

La population source est composée de 6 272 participants dont les caractéristiques sont présentées en Annexe 1. Pour être inclus dans la cohorte à l'étude, les participants devaient 1) être majeur, 2) avoir une infection primaire documentée avec une date et 3) avoir une infection secondaire (réinfection) documentée avec une date. Pour chaque individu, les données longitudinales disponibles pour l'étude couvraient une période maximale de trois ans entre le 18 mars 2020 et le 9 août 2023. La période de suivi de chaque participant pouvait alors varier, sans toutefois excéder trois années. Pendant cette période de suivi, le nombre d'événements était variable d'une observation à l'autre.

3.1.2 Données sur la population

La BQC19 collecte des données sur les participants, leurs habitudes de vie et leur état de santé ainsi que sur leur épisode de soins et les données cliniques afférentes. La liste complète des variables est disponible en Annexe 2. Les variables pertinentes pour le projet sont associées aux caractéristiques générales des individus comme leurs informations de nature sociodémographique et environnementale, mais également aux informations sur les événements de leurs épisodes de COVID-19 de même qu'à leurs caractéristiques vaccinales à leur sortie de l'étude.

D'abord, les variables sociodémographiques se rapportent aux participants de l'étude en soi telles que l'âge, le sexe à la naissance, le pays de naissance, etc. Certaines de ces variables sont plus spécifiquement axées sur le contexte de la COVID-19, notamment, deux variables indicatrices sur la profession, à savoir l'identification de statut de travailleur de la santé et celui de travailleur de laboratoire. Les variables sur l'environnement renseignent sur certains éléments de l'environnement des participants pertinents dans un contexte de contagion, notamment, deux variables catégorielles documentent respectivement le type de résidence du participant ainsi que la composition du ménage. Les variables événementielles décrivent les différents événements survenus au cours de l'hospitalisation (pour la cohorte maladie sévère) et des suivis longitudinaux subséquents à l'infection (pour les deux cohortes). Elles prennent différentes formes au sein du jeu de données, mais permettent généralement de déterminer l'événement et sa

date et, si applicable, d'obtenir de l'information afférente à l'événement. En tout, vingt variables renseignent sur la date et le résultat (positif ou négatif) de dix tests de dépistage du SRAS-CoV-2, trois variables documentent les dates des doses 1, 2 et 3 de vaccin et deux variables identifient partiellement les dates des réinfections. Ces différentes variables ont permis de créer les variables d'intérêt nécessaires au projet (voir section 3.3). Finalement, les variables sur l'état vaccinal renseignent sur le statut vaccinal du participant à sa sortie de l'étude (vacciné ou non) ainsi que sur le nombre de doses administrées.

3.2 Gestion des données

Dans une majorité de projets en sciences des données, l'étape de la préparation des données revêt une importance cruciale puisqu'elle génère et formate les intrants utilisés pour les analyses. Elle exerce ainsi une influence certaine sur la qualité des découvertes associées. Pyle (1999) estime que 70 % des efforts d'un projet de forage de données sont investis sur les données alors que Tufféry (2010) considère plutôt 38 %. Dans la pratique, les praticiens en sciences des données utilisent parfois la Loi de Pareto pour chiffrer que 80% des efforts d'un projet de forage de données sont investis sur les données. Quoi qu'il en soit, bien que la proportion d'effort ne fasse pas nécessairement consensus, il est unanime qu'elle revêt une grande importance. De ce fait, il s'avère pertinent d'y dédier une section dans le cadre de ce projet. Ainsi, dans cette section, la gestion et le traitement des données de l'étude seront abordés. Plus précisément, leur nettoyage ainsi que les différentes transformations appliquées pour les rendre optimales pour l'apprentissage automatique seront présentés. Cette étape du projet, réalisé pour l'ensemble de la population source, a notamment généré un script de nettoyage et de transformation autoportant pouvant être utilisé de manière systématique sur les jeux de données concernées permettant ainsi à la BQC19 d'en bénéficier et de le rendre disponible aux utilisateurs des jeux de données en question.

3.2.1 Nettoyage des données

Le nettoyage des données a été effectué par un processus itératif d'exploration des données et de traitement des éléments ainsi mis en évidence. Des croisements entre les

variables ont également été inclus dans les itérations afin de permettre des validations de cohérence entre les données corrélées ou reliées par le contexte. Les différents traitements réalisés se regroupent en trois types distincts à savoir la gestion des valeurs manquantes, la standardisation des valeurs et la validation croisée des variables.

3.2.1.1 Valeurs manquantes

Dans les jeux de données, la codification des valeurs manquantes varie entre les différentes variables. En effet, l'indication d'une valeur manquante prend plusieurs formes au sein du jeu de données soit 1) une valeur prévue à cet effet dans le dictionnaire des données, 2) un code HL7[3] ou 3) simplement l'absence d'une valeur. Ces différentes formes de valeurs manquantes étant différentes d'une variable à l'autre, et parfois même présentes simultanément sur une même variable, il était difficile d'envisager le nettoyage en lot des 1 034 variables sans risque sur l'intégrité des données comme le démontre la Figure 1. Un traitement variable par variable a plutôt dû être réalisé afin d'analyser le domaine de valeurs de la variable, de déterminer le contexte dudit domaine des valeurs et de recoder toutes les possibilités de valeur manquante vers la valeur standardisée.

A

hcworker	
Valeur	**NA**
0	Non
1	Non
..	..
..	..
..	..
99	Oui
NASK	Oui
UNK	Oui
..	..
..	..
..	..
..	..

B

temp_Route	
Valeur	**NA**
0	Oui
1	Non
2	Non
3	Non
4	Non
..	..
NASK	Oui
UNK	Oui
NAVU	Oui
NI	Oui
NM	Oui
< vide >	Oui

C

female	
Valeur	**NA**
0	Non
1	Non
2	Oui
..	..
..	..
..	..
..	..
..	..
..	..
..	..
..	..
..	..

Figure 1 - Exemples de variables dont le domaine des valeurs inclut différentes formes de valeurs manquantes (NA) : A) Travailleur de la santé, B) Site de prise de température et C) Sexe à la naissance. Les lignes en gris mettent en évidence, entre les trois exemples, une même valeur dont la signification de valeur manquante varie d'une variable à l'autre rendant impossible la recodification en lot des valeurs manquantes.

[3] Pour Health Level 7, un ensemble de standards internationaux d'interopérabilité pour l'échange, l'intégration, le partage et la récupération de données de dossier de santé électronique (Ait Abdelouahid et al., 2023). Ces standards prévoient notamment des valeurs prédéterminées selon le type de valeurs manquantes (p. ex. NASK - « *Not asked* », UNK - « *Unknown* », NI - « *No information* », etc.).

3.2.1.2 Standardisation des valeurs

Pour certaines variables, le domaine des valeurs n'était pas uniformisé, c'est-à-dire que pour une même modalité prise par la variable, plusieurs variations de la valeur existaient. Un traitement a été appliqué à chaque variable afin d'assurer que chaque valeur soit au final unique et que toutes les observations touchées pointent vers la valeur unifiée. La Figure 2 présente un exemple d'une telle situation avec un extrait d'une variable et d'une de ses modalités ainsi que le résultat suivant la standardisation.

A

birth_country (données brutes)	
Valeur*	**n**
Canada	3 782
CANADA	33
canada	4
Autres pays, Canada	3
Ontario	1
Not Canada	1
other country	203
Other country	2
NAVU	965
NI	4
UNK	201
Unknown	58
...	...

*128 valeurs distinctes

B

birth_country (données standardisées)	
Valeur**	**n**
Canada	3 820
NA	1 437
...	...

**110 valeurs distinctes

Figure 2 - Exemple de standardisation appliquée pour la variable du pays de naissance : données brutes (A) et données standardisées (B). L'exemple est un extrait de la variable du pays de naissance ciblé sur le Canada et les valeurs manquantes aux fins d'illustration. La variable présentait 128 et 110 valeurs distinctes avant et après le traitement respectivement.

De manière à standardiser la variable présentée à la Figure 2, la « *standard classification of countries and areas of interest (SCCAI) 2022* » (Statistique Canada, 2023) a été utilisée. Il s'agit d'un jeu de données publié en 2023 par Statistique Canada. Basé sur la norme ISO 3166-1:2020, ce dernier se veut une liste standardisée des pays et zones correspondant à l'usage général recommandé par le gouvernement canadien. Il comprend d'ailleurs les 249 éléments de la norme ISO 3166-1:2020 auxquels s'en ajoute deux, le pays du Kosovo et la zone Sercq, respectivement reconnue par le Canada comme un pays en 2008 et par les Nations Unies en 2011. Le nom standardisé du pays a permis la standardisation du lieu de naissance des participants.

3.2.1.3 Valeurs extrêmes, aberrantes et incohérentes

Plusieurs variables des jeux de données avaient des relations entre elles que ce soit d'un point de vue d'une corrélation ou de leur contexte. La nature de ces liens différait selon le contexte des variables et nécessitait ainsi une analyse et des tests spécifiques afin de

repérer les valeurs potentiellement extrêmes, aberrantes ou incohérentes. Les situations sont conséquemment diverses, mais trois principaux types de problématiques ont été identifiés.

Variables associées. Certaines variables associées entre elles prenaient des valeurs qui ne respectaient pas l'association. Par exemple, pour quelques observations, la variable indicatrice nourrisson (`infant`) prenait la valeur *VRAI* pour une variable âge (`age`) supérieure ou égale au maximum établi selon le dictionnaire des données pour être un nourrisson (< 1 an), considérant alors à tort des enfants de plus d'un an comme étant des nourrissons.

Tableau 1 - Exemple de l'association erronée entre la variable Nourrisson (infant) et la variable Âge (age)

Obs.	age (année)	Infant (initiale)	Infant (corrigée)
1	8,5	TRUE	FALSE
2	14,8	TRUE	FALSE
3	10,0	TRUE	FALSE

Variables temporellement associées. Dans un ordre d'idée similaire, l'incohérence entre certaines variables associées entre elles était plutôt longitudinale. Parmi les exemples, il est possible de citer la séquence des vaccins où les dates ne correspondaient pas à l'ordonnancement des doses (voir le Tableau 2) ou, encore, à une réinfection qui précédait l'infection.

Tableau 2 - Exemples d'incohérence temporelle impliquant la séquence des vaccins

	Séquence des vaccins				
Obs.	Vaccin_date1	<	Vaccin_date2	<	Vaccin_date3
1	2021-07-15	>	2021-04-17	<	2022-02-25
2	2021-01-12	<	2022-05-03	>	2021-12-17
3	2021-06-16	>	2021-03-13	<	2022-01-06

Contexte. Certaines discordances découlaient quant à elles de la valeur de la variable par rapport au contexte. Par exemple une hospitalisation avec un diagnostic d'admission de COVID-19 dont la date précédait le premier cas répertorié de SRAS-CoV-2 ou encore une date de vaccination COVID en 2011.

Vu la teneur des incohérences, le traitement requis dépassait les opérations de nettoyage usuel dans le sens où, bien que des hypothèses pouvaient être avancées pour expliquer la valeur erronée ou aberrante et l'imputer (ex. erreur de saisie de 2011 versus 2021 pour reprendre l'exemple ci-haut), les corrections ont été réalisées avec la collaboration de l'équipe de la BQC19 afin de confirmer la valeur réelle « corrigée ». Lorsque la valeur rectifiée n'était pas disponible, le patient était retiré de la cohorte (n =1).

3.2.2 Construction des jeux de données

Après nettoyage, les données ont été transformées de manière à être optimalement structurées pour les différentes analyses temporelles par apprentissage machine à réaliser et deux jeux de données principaux ont été créés. Le premier concerne les caractéristiques d'identification des participants de la cohorte et le deuxième recense les différents événements à l'étude (entrée dans la cohorte, infection, vaccination, réinfection…). D'autres jeux de données secondaires ont aussi été créés de manière à contenir les détails spécifiques à certaines observations du jeu de données des événements. Le schéma du processus de transformation globale est présenté en Annexe 3.

Des travaux de sélection et de transformation des variables ont été nécessaires afin de construire les jeux de données finaux à partir du jeu de données initial, puisque ce dernier comportait tous les types évènements confondus. À titre d'exemple, le statut vaccinal et le nombre de doses de vaccin des participants de l'étude n'avaient que des valeurs nulles dans l'événement d'identification du patient, mais contenaient des valeurs dans l'événement de sortie de l'étude. Pour faciliter le travail de repérage des variables, le jeu de données initial a été subdivisé en sous-jeu de données par type d'événement dans lesquels seules les variables dont au moins une observation contenait une valeur non nulle ont été conservées. Ces sous-jeux ont grandement facilité la compréhension des données par type d'événement et ont servi d'intrant à la création d'une table de *mapping* indiquant pour chaque variable de chacun des sous-jeux de données, leur destination dans les jeux de données finaux. La Figure 3 illustre la structure du jeu de données initial, l'étape de sélection et de repérage des variables ainsi que la constitution des jeux de données principaux et secondaires.

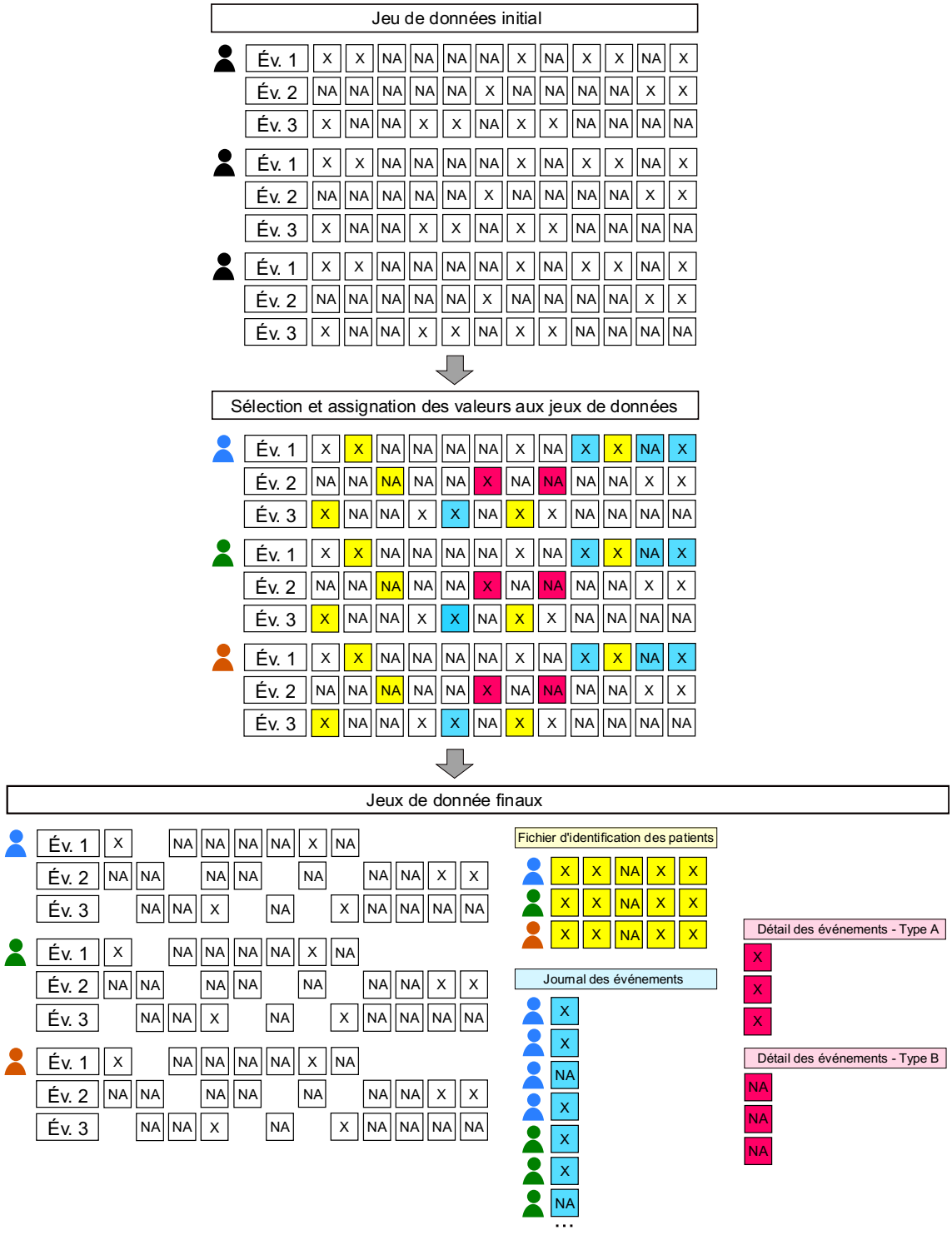


Figure 3 - Schéma de la sélection des variables pour la création de jeux des données finaux. Les éléments en jaune représentent les variables concernant les participants (ex. variables sociodémographiques), les éléments en bleus représentent les événements (ex. dates) et ceux en rose, les variables caractérisant l'événement (ex. résultat de l'événement « dépistage » : positif.

3.3 Variables d'intérêts

Dans le cadre de ce projet, nous ferons la distinction entre les variables d'intérêts utilisées pour générer les groupes par apprentissage automatique non supervisé de celles utilisées pour les caractériser. Le Tableau 3 résume leur utilisation.

3.3.1 Variables de regroupement

Afin de regrouper les individus en fonction de leur schéma d'infection / réinfection / vaccination, les variables d'intérêts se rapportaient principalement à la temporalité de ces événements. Tout d'abord, la date de l'infection primaire, inexistante dans le jeu de données initial, a été reconstruite à partir de la date du premier test PCR positif de chaque participant.

Concernant la variable de réinfection, celle-ci a dû être reconstruite puisque le jeu de données contenait deux variables présentant des informations discordantes (données manquantes ou différentes). Après discussion avec l'équipe de la BQC19, la variable a été redéfinit selon le schéma suivant : 1) la valeur des deux variables lorsqu'elles étaient identiques, 2) la valeur non manquante lors que l'une des variables était nulle et 3) la plus hâtive des dates lorsque les deux variables ne contenaient pas la même valeur. De plus, dans certains cas (N = 96), plusieurs événements de type réinfection étaient documentés. En concertation avec l'équipe de la BQC19, seule la première réinfection documentée pour chaque individu a été retenue systématiquement. Pour les observations subséquentes, lorsqu'elles ne n'étaient pas explicitement identifiées comme des suivis longitudinaux de la réinfection, elles ont été comparées aux autres. Pour être considérée comme une nouvelle réinfection, elle ne devait pas être à l'intérieur du délai interne fixé par la BQC19 à cet effet. Ainsi, si la date d'une observation ultérieure se situait à moins de 14 jours de la réinfection précédente, elle n'est pas identifiée comme une nouvelle réinfection.

Enfin, la date de vaccination a été extraite directement du jeu de données de manière à compléter les différents événements qui ont servi à établir la séquence temporelle d'infection, de réinfection et de vaccination de participants de la cohorte.

À partir de ces événements, des variables d'intérêts supplémentaires (N=21) ont été calculées de manière à obtenir le délai séparant chaque paire d'événements. Par exemple, cela signifie que pour l'infection initiale, des délais ont été calculés par rapport aux événements suivants : le premier, le deuxième et le troisième vaccin, ainsi que la première, la deuxième et la troisième réinfection. Une deuxième version normalisée de ces variables a été générée, c'est-à-dire une version avec une distribution centrée autour de zéro et avec une variance de 1. Cette normalisation dite centrée réduite a été obtenue en soustrayant la moyenne, puis en divisant par l'écart-type.

Tableau 3 - Résumé de l'utilisation des variables d'intérêt

Variables	**Utilisation**	
	Analyse de regroupement	**Caractérisation**
Date d'infection	•	
Date(s) de réinfection	•	
Date(s) de vaccination	•	
Délai interévénement	•	•
Âge (numérique et catégorie)		•
Indice de masse corporelle (numérique et catégorie)		•
Sévérité de la maladie		•
Consommation de drogue		•
Variant prédominant lors de l'infection		•
Variant prédominant lors de la réinfection		•
Vague de l'infection		•
Vague de la réinfection		•
Consommation de cigarette électronique		•
Sexe		•
Travailleur de la santé		•
Travailleur de laboratoire		•
Type de résidence		•
Ménage		•
Nombre de dose(s) de vaccin lors de l'infection		•
Nombre de dose(s) de vaccin lors de la réinfection		•
Nombre de réinfection(s)		•

3.3.2 Variables de caractérisation des groupes

Afin de décrire les différents groupes obtenus lors de l'analyse de regroupement basée uniquement sur les séquences temporelles, les variables d'intérêts concernent les caractéristiques sociodémographiques, l'état et les habitudes du participant ainsi que le contexte de la contagion. On y compte notamment l'âge du participant lors de l'entrée dans l'étude, un regroupement de l'âge selon la classification des catégories d'âge par

tranches de cinq ans de Statistique Canada (2022) en subdivisant toutefois la catégorie des 18 ans en moins pour isoler les nourrissons (< 1 an), la classe de l'indice de masse corporelle selon Santé Canada (2003), le sexe biologique et enfin le pays de naissance. Également, des variables concernant la consommation de tabac (non-fumeur, fumeur, ex-fumeur, tabagisme passif), de cigarette électronique (oui ou non) et de drogues (oui ou non) contribuent à brosser un portrait des habitudes de consommation des participants. De même, certaines informations sur l'environnement du patient s'avèrent pertinentes dans un contexte de transmission de la maladie pour caractériser les groupes dont des variables sur son lieu de résidence, sur les habitants de son ménage ainsi que son occupation (travailleur de la santé ou travailleur de laboratoire).

Ensuite, au chapitre des variables concernant le contexte des infections, notons le nombre de réinfection de chacun des individus, le nombre de doses de vaccin reçus lors des différents événements d'intérêts, à savoir lors de l'infection primaire et lors de chaque réinfection, ainsi que deux variables issues d'un calendrier interne à la BQC19 (voir Annexe 4), soit la vague pandémique et le variant prédominant. Il est toutefois important de mettre l'emphase sur le fait qu'il est question du variant prédominant dans la population et non du variant réel associé à l'infection ou la réinfection du participant. En effet, les données sur le séquençage des variants n'étaient disponibles que sur une infime partie de la cohorte (n = 20) et le variant réel n'a conséquemment pas été inclus dans les variables d'intérêts.

Finalement, les délais interévénements (mentionnés précédemment à titre de variables utilisées pour la génération des groupes) sont aussi des variables d'intérêts pour la description des groupes.

3.4 Analyses statistiques et visualisation

En premier lieu, des analyses descriptives ont été effectuées afin de décrire la cohorte. Notamment, la fréquence des effectifs de même que le mode ont été calculés pour les variables catégorielles alors que des mesures de tendance centrale (moyenne et médiane) et des mesures de dispersions (minimum, maximum, écart-type et variance) l'ont été pour les variables numériques.

Ensuite, ces mêmes analyses ont été reproduites, mais sur différente stratification de la cohorte (p. ex. par sexe, par profession, etc.). Des tests paramétriques de comparaisons de moyennes (test T de Student ou ANOVA) ou non paramétriques de rangs (Test de Wilcoxon ou Kruskal-Wallis) ont été aussi été effectués quand applicable. Lorsque les tests de comparaison multiples suggéraient une différence significative entre les groupes, des tests post-hoc supplémentaires, dont le test des étendues de Tukey et le test de Dunn, ont aussi été appliqués.

Pour finir, une analyse de fouille de processus (*« process mining »* en anglais), une famille d'analyses issues d'une combinaison de techniques de gestion de processus et de sciences des données (Munoz-Gama *et al.*, 2022), a été appliquée afin de visualiser les séquences d'événements de la cohorte grâce à la production de cartes présentant l'enchaînement temporel des événements.

3.5 Analyses de regroupement

Afin de grouper les individus selon la similarité de leur séquence temporelle (« *pattern* ») de vaccination, d'infection et de réinfection, des algorithmes d'apprentissage automatique non supervisés ont été utilisés afin de réaliser des analyses de regroupement. Plus spécifiquement, dans l'optique d'atteindre le sous-objectif d'exploration des méthodes d'apprentissage automatique en respectant un équilibre performance / interprétabilité, différentes combinaisons d'algorithmes, de mesures de dissimilarité et de données ont été testés. La Figure 4 résume les différentes combinaisons et présente également la structure des différentes sources de données. Les combinaisons retenues ont ensuite été évaluées en termes de performance, de stabilité et de sensibilité.

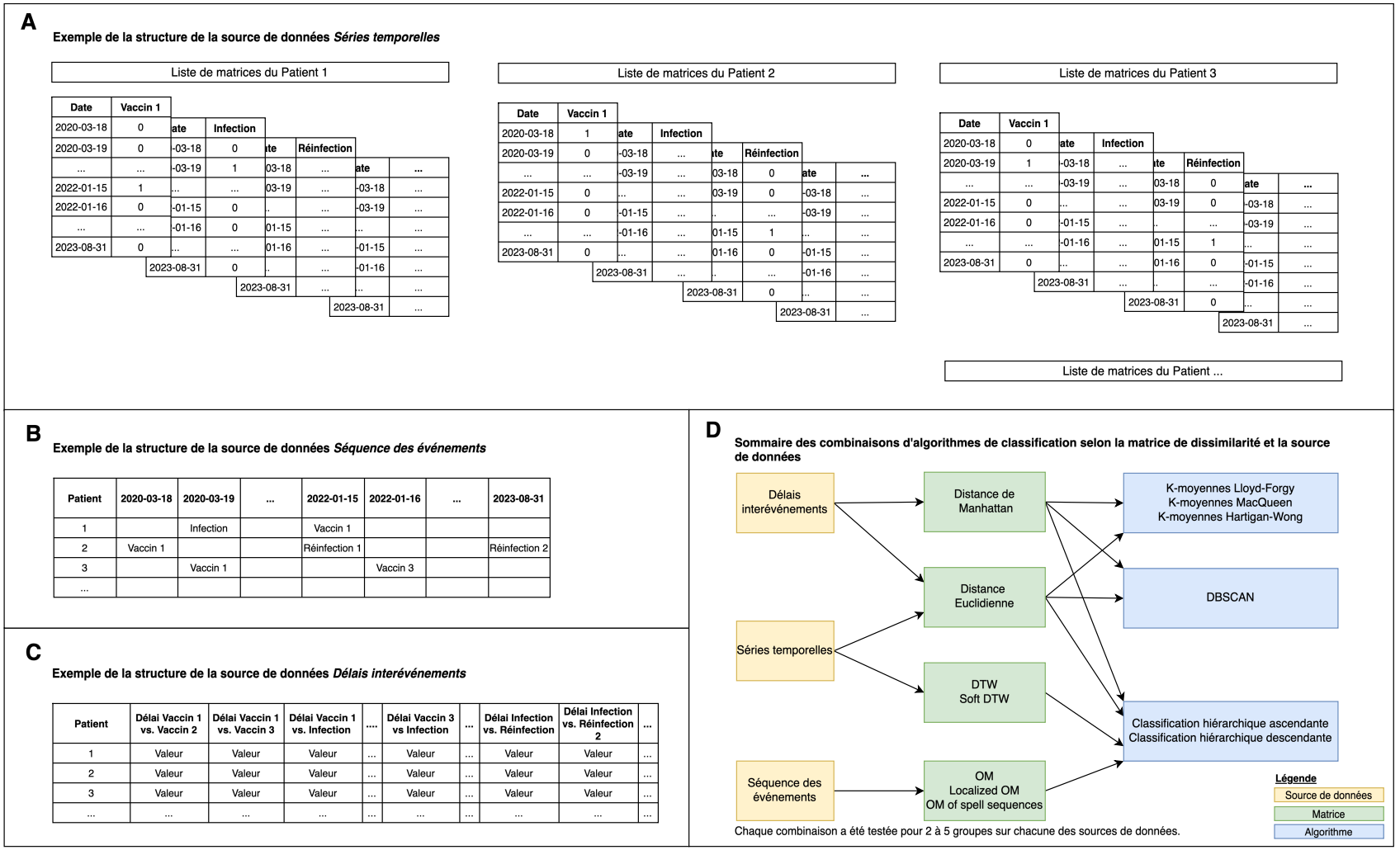


Figure 4 - Sommaire des combinaisons et structures des sources de données où A) est la structure pour les séries temporelles, B) la structure pour les séquences des événements, C) celle pour les délais interévénements et D) le sommaire des différentes combinaisons d'algorithmes de classification, de matrice de dissimilarité et de sources de données.

3.5.1 Mesure de dissimilarité

Une mesure de dissimilarité est une métrique qui calcule, à partir d'une méthode donnée, à quel point deux observations sont différentes. Elles sont généralement insérées dans une matrice appelé matrice de dissimilarité qui donne, paire par paire, le résultat de cette mesure explicitant ainsi la différence entre chaque observation. Plusieurs méthodes de calcul de la dissimilarité existent, telles que les méthodes classiques de distance Euclidienne ou la distance de Manhattan, ou bien des méthodes tenant compte des séries temporelles telles que l'*Optimal matching* (OM) et le *Dynamic time warping* (DTW). Ces méthodes, ayant été utilisées dans ce projet, seront présentées ci-après. Ces différentes matrices de dissimilarité ont servi d'intrant aux algorithmes de classification afin de générer les différents groupes en fonction de la similarité ou dissimilarité de leur séquence temporelle.

3.5.1.1 Distance Euclidienne et distance de Manhattan

La distance Euclidienne correspond à la longueur d'un segment de droite qui relie deux observations dans un plan cartésien à deux dimensions. La distance de Manhattan, quant à elle, correspondant à la distance entre deux observations, mais en ne réalisant que des angles droits. Ainsi, si l'on référait au théorème de Pythagore, la distance Euclidienne est l'hypoténuse alors que la distance de Manhattan correspond à l'addition des deux cathètes du triangle.

3.5.1.2 Dynamic time warping

Le *Dynamic time warping* (DTW) permet de mesurer la similarité entre des séquences temporelles et de trouver l'alignement optimal moyennant la prise en considération des différentes déformations dans les séries (séquences, délais entre les événements, répétition, etc.) (Sakoe et al., 1978). Pour le projet, l'algorithme a été utilisé dans un contexte multidimensionnel où, pour chaque individu de la cohorte, une série temporelle distincte pour chaque type d'événement d'intérêt (infection primaire, vaccination et réinfection) était prise en compte simultanément par l'algorithme. Ce dernier cherche alors à trouver l'alignement optimal, en tenant compte des trois séries temporelles, qui minimise la distance totale par rapport aux séries des autres individus. La distance totale

de l'alignement optimal ainsi obtenu permet de générer une matrice de dissimilarité entre chaque paire d'individu.

Trois versions du DTW ont été appliquées. La première correspondait à la version originale de l'algorithme permettant de déterminer l'alignement « optimal ». La seconde variante utilisait la norme Euclidienne pour construire la matrice de distance (Sardá-Espinosa, 2017). La troisième, dite *soft-DTW* (Cuturi et Blondel, 2017), calculait la somme pondérée de tous les alignements possibles afin de rendre la fonction plus lisse et robuste aux variations mineures (Cuturi *et al.*, 2017).

3.5.1.3 Optimal matching

L'*optimal matching* (OM) a pour but la comparaison de séquences complexes entre elles et permet notamment d'en extraire les « *patterns »* afin de les analyser (Abbott et Forrest, 1986; Abbott et Hrycak, 1990). L'algorithme mesure notamment la similarité ou la dissimilarité des séquences entre elles en minimisant le coût attribué aux opérations appliquées aux séquences pour les rendre similaires (insertion, suppression ou substitution d'événement). Selon la nature du projet, il n'était pas désiré de discriminer une de ces opérations plus qu'une autre impliquant que le coût a été établi selon la méthode constante en laissant la valeur par défaut de 2. À partir de cette matrice de coûts de substitution, l'algorithme OM a été utilisé pour calculer une matrice de dissimilarité entre les différentes séquences d'événements des participants de la cohorte.

Deux variantes supplémentaires de *l'optimal matching* ont également été utilisées, soit la version localisée (*localized OM*) et la version par période (*OM of spell sequences*) (Studer et Ritschard, 2015, 2016). La première variante permet de focaliser l'analyse sur des segments plutôt que sur la totalité de la séquence d'événements des individus alors que la seconde l'est plutôt sur la durée des événements de la séquence de chaque individu. Ces deux variantes ont été testées dans le contexte de l'approche basée sur les données (*data driven*), sachant à priori que leur utilisation n'est pas nécessairement optimale avec les données disponibles.

3.5.2 Algorithmes de classification

Cette section présente les différents algorithmes de classification qui ont été appliqués sur les matrices de dissimilarité pour la génération des groupes en fonction de leur séquence temporelle. À titre de rappel, le choix des algorithmes repose, au même titre que les matrices de dissimilarité, sur une volonté d'interprétabilité et d'adéquation avec les données.

3.5.2.1 Classification hiérarchique

Deux types d'analyses de classification hiérarchique, soit ascendante (agglomérative) et descendante (divisive), ont été réalisées dans le cadre du projet. La première méthode présuppose que, initialement, chaque observation est un groupe qui, à chaque itération, se voit fusionné avec le groupe le plus près, et ce, jusqu'à ce que l'ensemble des observations forment un seul et même groupe. Il existe différentes façons de former les groupes lors des itérations qui, dans le cadre du projet, ont été selon les plus proches voisins (*simple linkage*), selon les voisins les plus éloignés (*complete linkage*), avec la moyenne des distances entre les paires des observations inter-groupes (*average linkage*), avec le barycentre des groupes (*centroid linkage*) et selon la minimisation de l'homogénéité globale (*Ward*). La deuxième méthode, la classification hiérarchique divisive, utilise la logique inverse de l'agglomérative en partant d'un groupe qui contient toutes les observations, pour ensuite le subdiviser, à chaque itération, jusqu'à l'obtention d'un groupe pour chaque observation.

3.5.2.2 K-moyennes

Des analyses basées sur la méthode k-moyennes (« *k-means* » en anglais) ont également été menées, plus précisément avec trois variantes, soit Lloyd-Forgy (Forgy, 1965; Lloyd, 1982), MacQueen (1967) et Hartigan-Wong (1979).

Dans le cas de la méthode k-moyennes de Lloyd-Forgy, un point central, le centroïde, est initialisé de façon aléatoire pour chaque groupe déterminé a priori. Chaque observation est alors assignée au centroïde le plus près selon la distance Euclidienne afin de former un groupe. Le centroïde est ensuite déplacé pour correspondre à la moyenne, ou encore le centre du groupe nouvellement formé. Ces deux étapes sont répétées jusqu'à l'atteinte

d'une convergence, c'est-à-dire jusqu'à ce qu'il n'y ait plus de changement significatif dans le processus itératif d'assignation des observations / déplacement du centroïde.

Le principe de la méthode k-moyennes de MacQueen est similaire à la variante Lloyd-Forgy, à l'exception du déplacement du centroïde. En effet, le centroïde est déplacé à chaque nouvelle assignation d'une observation au groupe plutôt qu'après l'assignation de toutes les observations.

Enfin, la variante Hartigan-Wong utilise une logique itérative similaire à MacQueen, en poursuivant toutefois l'objectif de diminuer au maximum la variance intragroupe à chaque nouvelle assignation d'une observation au groupe.

3.5.2.3 Regroupement spatial basé sur la densité avec bruit

Des analyses basées sur le modèle « *density-based spatial clustering of applications with noise » (DBSCAN) (Ester et al., 1996)* ont également été réalisées. Pour générer les groupes, l'algorithme se base sur les plus proches voisins d'un point. Si ce dernier dispose d'un nombre suffisant de voisins, il devient le point central d'un groupe. Les deux principaux hyperparamètres de l'algorithme sont l'epsilon (Eps), à savoir le rayon pour qu'une observation soit considérée comme voisin d'un autre, et le minimum de points (MinPts), c'est-à-dire le nombre d'observations requises pour qu'un groupe soit constitué. Ainsi, pour un point donné, si le MinPts est rencontré à l'intérieur de l'Eps, il formera le centroïde d'un groupe. Ce processus est répété jusqu'à ce que tous les points aient été analysés.

Le paramètre MinPts devrait être égal au minimum du nombre de dimensions du jeu de données selon l'article original. Sander *et al.* (1998) estiment cependant qu'il devrait correspondre au double du nombre de dimensions lorsque plus de deux dimensions sont impliquées. À cet effet, le MinPts a été fixé à 34 (2*17 dimensions). Afin de déterminer visuellement la valeur optimale de Eps pour ce paramètre, un graphique de la distance des k plus proches voisins (« *k-nearest neighbors distance »*) a été généré où k prenait la valeur de MinPts – 1. Une augmentation soudaine suggérant que les valeurs à droite étaient des valeurs aberrantes, la valeur juste avant ce coude était celle retenue pour l'Eps.

3.5.3 Approches de regroupement

Considérant l'approche basée sur les données (« *data-driven* ») adoptée pour le projet, plusieurs combinaisons de sources de données, de mesures de dissimilarité et d'algorithmes ont été explorées afin d'évaluer les résultats obtenus. Ces différentes combinaisons ou approches seront présentées en fonction de la source de données sur laquelle reposait l'analyse, c'est-à-dire la séquence temporelle complète des évènements, ou bien les délais entre ces évènements. L'ensemble des combinaisons ont été testées pour un nombre variant de 2 à 5 groupes. Afin de discriminer les différents modèles de regroupements obtenus, une mesure de performance, le coefficient de silhouette (Rousseeuw, 1987), a guidé la sélection des modèles issus des différentes combinaisons. Variant de -1 à 1, la moyenne du coefficient des observations permet de mesurer la cohérence de leur assignation à leur groupe par rapport à la dissimilarité aux observations des autres groupes, orientant ainsi l'analyse de la qualité des regroupements. En surplus, le nombre d'individus dans chacun des groupes a également été considéré afin de s'assurer qu'il était suffisant pour des interprétations acceptables.

3.5.3.1 Partitionnements basés sur la séquence des événements

Les dates des événements ont servi de sources de données à deux approches, et ce, selon deux niveaux de granularité de données, à savoir 1) au niveau du jour de l'événement et 2) au niveau de la semaine de l'événement.

Ainsi, une première approche exploitait, à l'aide des différentes versions du DTW présentées précédemment, les événements sous forme de séries temporelles. La Figure 4A démontre un exemple de la structure de données pour cette approche. Les multiples combinaisons de matrices de dissimilarité alors obtenues ont servi d'intrant à des classifications hiérarchiques ascendantes et descendantes. En tout, 20 associations de paramètres différentes ont donc été testées sur les deux granularités de données pour un total de 160 modèles différents. Les résultats présentés dans l'article du Chapitre 4 découle de cette approche.

Une deuxième approche employait les événements sous forme de séquences pour établir, à l'aide de l'OM et ses variantes, des matrices de dissimilarité. Un exemple de la structure des données utilisées avec cette approche est présenté à la Figure 4B. Des classifications hiérarchiques ascendantes et descendantes ont été réalisées sur ces différentes combinaisons des matrices. En tout, 21 associations de paramètres différentes ont donc été testées sur les deux granularités de données mentionnées plus tôt pour un total de 168 modèles de regroupements différents.

3.5.3.2 Partitionnements basés sur les délais interévénements

Les variables de délais entre les paires d'événements ont été utilisées comme intrant à trois approches supplémentaires, et ce, en versions normalisée et non-normalisée. Un exemple de la source de données est présenté à la Figure 4C.

La troisième approche constituait en des classifications hiérarchiques ascendante et descendante dont les différentes combinaisons de méthodes de génération des groupes ont été testées sur des matrices de distance calculées avec la distance Euclidienne et la distance de Manhattan. Ainsi, 88 modèles de regroupement ont été générés.

Une quatrième approche a été réalisée dans un esprit similaire avec les trois variantes de l'algorithme k-moyennes résultant en 24 modèles de regroupements supplémentaires.

La cinquième et dernière approche était basée sur l'algorithme du regroupement spatial basé sur la densité avec bruit (DBSCAN). L'algorithme étant sensible au choix des hyperparamètres, des combinaisons additionnelles à celles issues de la littérature et présentées plus tôt ont été testées afin d'ouvrir un spectre plus large : MinPts = [25, 50], Eps = [0.15, 90, 170]. 36 modèles de regroupements ont dans cet optique été créés.

3.5.4 Caractérisation des groupes

Afin de caractériser les groupes obtenus à partir de l'analyse de regroupement, des analyses descriptives ont été réalisées. Ces dernières, appliquées sur les variables présentées à la section 3.3.2, sont essentiellement les mêmes analyses statistiques que

celles précédemment discutées à la section 3.4, mais sur les partitionnements résultants des analyses de regroupement.

3.5.5 Analyses de stabilité et de sensibilité

Considérant le nombre de méthodes différentes expérimentées, des analyses de sensibilité ont été produites de manière à déterminer le rôle de l'algorithme et de ses paramètres dans les résultats et, conséquemment, évaluer la robustesse des analyses et des résultats afférents.

Tout d'abord, un modèle a été sélectionné sur la base du coefficient de silhouette le plus élevé pour les méthodes du DTW, du OM et des classifications hiérarchiques ascendante et descendante sur les délais interévénement. Lorsqu'au sein d'une même méthode, plusieurs modèles étaient ex aequo, le modèle ayant les coefficients de silhouette les plus élevés pour les regroupements adjacents (k-1 et k+1) a été retenu. Les analyses de sensibilité ont été effectuées sur ces modèles.

Constance des attributs distinctifs. À partir des analyses descriptives des groupes, les caractéristiques ou les attributs mis en évidence ont été comparés pour chaque modèle de regroupement sélectionné dans l'objectif valider si 1) un élément d'intérêt mis en lumière dans un modèle est également présent dans les autres ou 2) les éléments d'intérêt mis de l'avant dans les autres modèles sont absents d'un modèle. Sachant que le nombre optimal de groupes peut varier d'une méthode à l'autre, le but n'était pas de rechercher pour un groupe donné son homologue dans une autre méthode, mais plutôt de vérifier entre les méthodes une certaine constance dans ce que les statistiques descriptives des groupes révèlent.

Stabilité des groupes. La stabilité des clusters en termes de cohérence d'affectation a été analysée. À cet effet, pour chaque modèle sélectionné, l'assignation de chaque paire d'individus au même groupe a été validée. Ainsi, quatre matrices ont été générées où l'intersection entre deux individus prenait la valeur *1* lorsque la paire de participants était assignée au même groupe. Ces matrices ont ensuite été additionnées permettant de mettre en lumière au travers des différentes méthodes la fréquence où les individus ont été regroupés au sein du même groupe par les différents algorithmes.

3.6 Considérations éthiques

Le mémoire s'inscrivant dans un projet en collaboration avec le Dr. Simon Rousseau de l'Institut de recherche du Centre universitaire de santé McGill (CUSM), le comité d'éthique de la recherche (CER) évaluateur est celui rattaché au projet principal *Determining the impact of hybrid immunity on the evolving landscade of host responses to SARS-CoV-2 in the Biobanque Québécoise de la COVID-19 (BQC19),* c'est-à-dire celui du CUSM (réf. 2023-9261). Le CER de HEC Montréal, en vertu de *l'Entente pour la reconnaissance des certificats d'éthique des projets de recherche à risque minimal entre les universités*, a émis, pour le mémoire, une reconnaissance de l'approbation éthique du CER du CUSM (réf. 2024-5571).

3.7 Logiciels utilisés

Le nettoyage de données ainsi que les analyses ont été effectués avec les logiciel *R*, version 4.3.0 (R Core Team, 2023), *Rstudio,* version 2023.12.1.402 (Posit team, 2024) ainsi que la librairie *tidyverse*, version 2.0.0 (Wickham *et al.*, 2019). Plus spécifiquement, l'apprentissage machine a également été réalisée à l'aide des librairies *cluster,* version 2.1.6 (Maechler *et al.*, 2023), *factoextra*, version 1.0.7 (Kassambara et Mundt, 2020), *fpc*, version 2.2-11 (Hennig, 2023), *mclust,* version 6.1 (Scrucca *et al.*, 2023), *weightedcluster*, version 1.6.4 (Studer, 2013) pour les analyses de regroupement classiques, la librairie *TraMineR* version 2.2-9 (Gabadinho *et al.*, 2011; Studer *et al.*, 2016) pour les analyses impliquant l'algorithme Optimal matching, les librairies *dtw*, version 1.23-1 (Giorgino, 2009), *dwtclust*, version 5.5.12 (Sarda-Espinosa, 2023), *proxy,* version 0.4-27 *(Meyer et Buchta, 2022), ggseqplot,* version 0.8.3*, (Raab M, 2022)* pour les analyses impliquant l'algorithme Dynamic time warping et les librairies *jmv*, version 2.5.6 (Selker *et al.*, 2024), *rstatix*, version 0.7.2 (Kassambara, 2023), *car*, version 3.1-2 (Fox et Weisberg, 2019), *pheatmap,* version 1.0.12 (Kolde, 2019) et *bupaR,* version 0.5.4 (Janssenswillen *et al.*, 2019) pour la caractérisation des groupes et les analyses de sensibilité.

Chapitre 4 Article

Note : La mise en forme de ce chapitre respecte celle de l'éditeur de la revue et diffère conséquemment du reste du document.

CHARACTERIZATION OF HYBRID IMMUNITY TO SARS-COV-2 FROM THE BIOBANQUE QUÉBÉCOISE DE LA COVID-19 (BQC19)

AUTHORS

Author	Affiliation	ORCID
Jean-Frédéric Boulianne	HEC Montréal, Montréal, Québec, Canada Centre de recherche en santé publique (CReSP), Montréal, Québec, Canada	0009-0009-2717-9614
Denis Larocque	HEC Montréal, Montréal, Québec, Canada	0000-0002-7372-7943
Simon Rousseau*	McGill University, Montréal, Québec, Canada	0000-0002-8773-575X
Delphine Bosson-Rieutort*	École de santé publique de l'Université de Montréal, Montréal, Québec, Canada Centre de recherche en santé publique (CReSP), Montréal, Québec, Canada	0000-0002-5035-6457

* These authors contributed equally

Corresponding author: Jean-Frédéric Boulianne
jean-frederic.boulianne@hec.ca

Conflict of interest: The authors declare no competing interests.

Acknowledgement: This work was made possible through open sharing of data and samples from the Biobanque québécoise de la COVID-19, funded by the Fonds de recherche du Québec - Santé, Génome Québec, the Public Health Agency of Canada and, as of March 2022, the ministère de la Santé et des Services sociaux. We thank all participants to BQC19 for their contribution.

https://www.quebeccovidbiobank.ca

Characterization of hybrid immunity to SARS-CoV-2 in the Biobanque québécoise de la COVID-19 (BQC19)

Abstract

Objectives This study aimed to group individuals according to their pattern of infection, reinfection and vaccination sequences in the context of COVID-19 and to characterize these groups regarding socio-demographic, clinical and temporal factors.

Methods We applied machine learning methods to perform cluster analysis on a cohort of patients from the Biobanque Québécoise de la COVID-19. More specifically, we performed agglomerative and divisive hierarchical clustering on a distance matrix computed with the Dynamic time warping algorithm on the time series of patients' COVID-19 episodes.

Results The cohort participants were grouped into five clusters, and their characterization revealed that the clusters follow a temporal progression according to the timing of infection and its positioning across the waves of the pandemic. Reinfections, on the other hand, occurred from the fifth wave onwards. The most highly vaccinated groups appear to have been infected, and consequently reinfected, later in the pandemic. Some groups featured a higher proportion of healthcare workers, while for others, it was the trajectory and their timeframes that were of interest.

Keywords *SARS-CoV-2, COVID-19, immunity, cluster analysis, BQC19*

Résumé

Objectifs Cette étude avait pour objectifs de regrouper les individus selon leur similarité de séquences d'infection, de réinfection et de vaccination dans un contexte de COVID-19 et de caractériser ces groupes en termes de facteurs sociodémographiques, cliniques et temporels.

Méthodes Nous avons appliqué des méthodes d'apprentissage automatique pour réaliser des analyses de regroupement sur une cohorte de patients issus de la Biobanque québécoise de la COVID-19. Plus précisément, nous avons réalisé des classifications hiérarchiques ascendantes et descendantes sur une matrice de distance calculées avec l'algorithme Dynamic time warping sur les séries temporelles des épisodes de COVID-19 des patients.

Résultats Les participants ont été regroupés en cinq groupes et la caractérisation a révélé que ces derniers suivent une progression temporelle par rapport au moment de l'infection dans les différentes vagues de la pandémie. Les réinfections sont quant à elles survenues plus vers la fin, soit à compter de la cinquième vague. Les groupes les plus vaccinés apparaissent comme étant infectés, et conséquemment réinfectés, plus tard dans la pandémie. Certains

groupes mettent en lumière une proportion plus importante de travailleurs de la santé alors que pour d'autres, c'est plutôt les trajectoires et les délais qui étaient d'intérêt.

Mots-clés *SARS-CoV-2, COVID-19, immunité, analyse de regroupement, BQC19*

1.1 Introduction

In December 2019, a novel coronavirus (SARS-CoV-2) was identified in Wuhan, China. Few weeks after, in January 2020, World Health Organization (WHO) confirms an interhuman transmission. Rapidly, the virus spread across the planet, leading to the COVID-19 pandemic. By April 2024, this pandemic had resulted in over 7 million deaths and 770 million infections worldwide (Livieratos, Gogos et Akinosoglou, 2024; World Health Organization, 2024), making it the most significant pandemic since the 1918 Spanish flu (Lapuente, Winkler et Tenbusch, 2024).

Coronavirus can infect different animals and causes, in humans, moderate to severe respiratory infections. Although most symptoms resembled those of a typical respiratory disease, including fever, fatigue and cough (Hu et al., 2021; Zhu et al., 2020), a significant proportion of infections progressed to a more severe, even critical, form of the disease, potentially involving dyspnea, acute respiratory distress syndrome and multiple organ failure. In addition to the acute form of the disease, the infection could lead to persistent health problems (e.g. fatigue, shortness of breath, cognitive problems) known as post-COVID-19 syndrome, also known as long COVID (Lapuente et al., 2024; World Health Organization and Kryuchkov, 2022)

The pandemic revealed that individuals who contracted the disease were not permanently immunized and could be reinfected after a relatively short period of 7 to 12 months (Misra et Theel, 2022; Rodriguez Velásquez et al., 2024). This variability in reinfection time can be attributed to individual sensitivities, infection severity, and the evolution of the SARS-CoV-2 virus through various variants. These various mutations generate different immune responses that impact the immunity of individuals who have contracted the disease.

In this context, we aim to study hybrid immunity, the immunity conferred by a combination of infection, reinfection and vaccination, by identifying and characterizing SARS-CoV-2 reinfection profiles. To specifically tackle this temporal and complex aspect of the hybrid immunity, we propose to use machine learning techniques on data from *Biobanque québécoise de la COVID-19* (BQC19) to group individuals according to their temporal pattern of vaccination, infection, and reinfection only and then, characterize the groups thus obtained in terms of sociodemographic and clinical factors to highlight any hidden patterns or characteristics that could lead to similar temporal sequences across our population.

1.2 Methods

BQC19 is a multicenter database involving a network of 11 Quebec hospitals with five partner academic institutions and has been described elsewhere (Tremblay et al., 2021). Briefly, this pan-provincial initiative collects, stores and shares data and blood samples from COVID-19 patients, both severe and non-severe cases. The biobank contains several datasets about participants' characteristics, events and certain biological data (RNA, DNA, serum, plasma and peripheral blood mononuclear cells, etc.). It also includes longitudinal follow-up for 24 months following hospitalization (inpatient) or PCR testing (outpatient). The source population consists of 6,272 participants included between March 2020 and August 2023. To be included in the study cohort, participants had to 1) have reached 18 years old, 2) have a documented primary infection with a date, and 3) have a documented secondary infection (reinfection) with a date. For each individual, the longitudinal data available for the study ranged between two months and three years. As each participant's follow-up period may vary during the study, the number of events differed from one individual to another.

Data management was performed through an iterative process of data exploration and processing of the highlighted elements. Data exploration included frequency and mode for categorial variables, measures of central tendency (mean and median) and measures of dispersion (minimum, maximum, standard-deviation (SD) and variance) for numerical variables. Stratified analyses were also performed (e.g. sex, occupation, etc.) with parametric (t-test or ANOVA) or non-parametric (Wilcoxon or Kruskal-Wallis) statistical test as required. Post-hoc tests (Tukey's range test or Dunn’s test), when appropriate, were also conducted. Data management iterations also included

standardization of values domain and cross-validation between variables to enable consistency validations between data correlated or linked by context. The various BQC19 data collections were merged to reduce missing values in the main dataset, mostly for events details. Process mining analysis was also used, mainly to help visualization of events sequences and flows (Janssenswillen et al., 2019).

To achieve the aim of grouping individuals according to their specific events patterns, variables of interest about infection, reinfection and vaccination were used for the clustering. Firstly, as the date of primary infection did not exist in the initial dataset, it was reconstructed from the date of each participant's first positive PCR test. Secondly, the reinfection variable had to be reconstructed, since the dataset contained two variables with conflicting information (missing or different data). The variable was redefined as follow: 1) the value of the two variables when they were identical, 2) the non-missing value when one of the variables was null, and 3) the earliest of the dates when the two variables did not contain the same value. In some cases (N = 96), more than one reinfection event was documented and only the first documented reinfection for each individual was systematically retained. When subsequent observations were not explicitly identified as longitudinal reinfection follow-ups, they were compared with the others available dates and was not considered as a new reinfection if the date of a subsequent observation was within 14 days of the previous reinfection. This period corresponds to the internal delay set by BQC19 to consider a reinfection. Finally, vaccination dates were directly extracted from the dataset.

Participants were grouped using an agglomerative hierarchical clustering analysis using Ward's minimum variance method on a multidimensional times series dynamic time warping (DTW) dissimilarity matrix constructed using the temporal variables of interest. Dynamic time warping has been chosen because it was the most suitable method for the type of data, especially as it takes into account temporal deformations in order to align sequences (Sakoe and Chiba, 1978). The number of groups was determined following the average Silhouette statistic (Rousseeuw, 1987). For each group, a process map has been generated to support the visualization of their specific temporal sequence of events (infections, vaccinations and reinfections) (Giorgino, 2009).

To describe the different groups obtained from the cluster analysis based on the temporal sequence and highlight potential characteristics leading to similar sequences, we used variables related to socio-demographic characteristics (e.g. age, sex at birth, BMI), participant's condition, habits and environment (e.g. smoking and drug use status, occupation, household information) and the context of the contagion (e.g. number of reinfection, number of vaccine received when infection or reinfection occurs, pandemic's wave, predominant variant). However, as variant sequencing data were only available for a small proportion of the cohort (n = 20), the actual variant was consequently not included in the variables of interest. Finally, delays between each pair of events were calculated and used to describe each cluster. All statistical analyses previously mentioned were used to describe each variable stratified by cluster.

Data cleaning and analyses were performed using R, version 4.3.0 (R Core Team, 2023) with *Rstudio,* version 2023.12.1.402 (Posit team, 2024) supported by package *tidyverse*, version 2.0.0 (Wickham et al., 2019). Hierarchical clustering and dynamic time warping was performed using packages *dtw*, version 1.23-1 (Giorgino, 2009), *dwtclust*, version 5.5.12 (Sarda-Espinosa, 2023) and *proxy,* version 0.4-27 *(Meyer et Buchta, 2022).* Process mining for events visualization was performed using package *bupaR*, version 0.5.4 (Janssenswillen et al., 2019)

1.3 Results

Among 6,272 individuals, 318 participants with at least one reinfection were included in the study, with an average age of 43 (SD 13.8). Table 1 reports characteristics of individuals in the study cohort. Among them, 230 were women (72.3%) and 141 were healthcare workers (44.3%). A total of 31 participants were reinfected twice (9.3%), including 6 who were reinfected three times (1.9%). The average dose of vaccine at primary infection was 1.08 (SD 1.30) and average dose of vaccine at reinfections were respectively 2.36 (SD 0.876), 2.29 (SD 0.864), 2.50 (SD 0.548).

Table 1 – Sociodemographic and temporal characteristics of the study population between 2020/03 and 2023/08, and clusters

Characteristics	**Cohort (n = 318)**	**Clusters**				
		1 (n = 138)	**2 (n = 42)**	**3 (n = 11)**	**4 (n = 51)**	**5 n = 76)**
Age years: mean (SD)	43.0 (13.8)	43.2 (14.0)	40.5 (12.3)	41.8 (16.5)	40.2 (11.8)	46.3 (14.6)
18 to 34 *n*(%)	94 (29.6%)	42 (30,4%)	16 (38.1%)	4 (36.4%)	16 (31.4%)	16 (21.1%)
35 to 44 *n*(%)	84 (26.4%)	38 (27.5%)	7 (16.7%)	1 (9.1%)	20 (39.2%)	18 (23.7%)
45 à 64 *n*(%)	125 (39.3%)	51 (37.0%)	19 (45.2%)	5 (45.5%)	14 (27.5%)	36 (47.4%)

Characteristics	Cohort (n = 318)	Clusters				
		1 (n = 138)	2 (n = 42)	3 (n = 11)	4 (n = 51)	5 n = 76)
+ 65 years *n*(%)	15 (4.7%)	7 (5.1%)	0 (0%)	1 (9.1%)	1 (2.0%)	6 (7.9%)
BMI: mean (SD)	26.9 (5.32)	28.9 (8.62)	26.5 (5.27)	27.0 (5.86)	28.0 (5.14)	26.9 (5.32)
Sex at birth						
Female *n*(%)	230 (72.3%)	97 (70.3%)	26 (61.9%)	7 (63.6%)	41 (80.4%)	59 (77.6%)
Male *n*(%)	88 (27.7%)	41 (29.7%)	16 (38.1%)	4 (36,4%)	10 (19.6%)	17 (22.4%)
Smoking status						
Non-smoker *n*(%)	230 (72.3%)	101 (73.2%)	26 (61.9%)	5 (45.5%)	40 (78.4%)	58 (76.3%)
Smoker *n*(%)	22 (6.9%)	9 (6.5%)	8 (19.0%)	0 (0%)	2 (3.9%)	3 (3.9%)
Former smoker *n*(%)	55 (17.3%)	20 (14.5%)	6 (14.3%)	5 (45.5%)	9 (17.6%)	15 (19.7%)
Passive smoker *n*(%)	1 (0.3%)	1 (0.7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Missing *n*(%)	10 (3.1%)	7 (5.1%)	2 (4.8%)	1 (9.1%)	0 (0%)	0 (0%)
Use electronic cigarettes						
Yes *n*(%)	13 (4.1%)	3 (2.2%)	4 (9.5%)	1 (9.1%)	3 (5.9%)	2 (2.6%)
No *n*(%)	299 (94.0%)	131 (94.9%)	37 (88.1%)	9 (81.8%)	48 (94.1%)	74 (97.4%)
Missing *n*(%)	6 (1.9%)	4 (2.9%)	1 (2.4%)	1 (9.1%)	0 (0%)	0 (0%)
Use Drugs						
Yes *n*(%)	30 (9.4%)	10 (7.2%)	10 (23.8%)	0 (0%)	5 (9.8%)	5 (6.6%)
No *n*(%)	282 (88.7%)	124 (89.9%)	31 (73.8%)	10 (90.9%)	46 (90.2%)	71 (93.4%)
Missing *n*(%)	6 (1.9%)	4 (2.9%)	1 (2.4%)	1 (9.1%)	0 (0%)	0 (0%)
Healthcare worker						
Yes *n*(%)	141 (44.3%)	44 (31.9%)	9 (21.4%)	4 (36.4%)	26 (51.0%)	58 (76.3%)
No *n*(%)	171 (53.8%)	91 (65.9%)	31 (73.8%)	7 (63.6%)	24 (47.1%)	18 (23.7%)
Missing *n*(%)	6 (1.9%)	3 (2.2%)	2 (4.8%)	0 (0%)	1 (2.0%)	0 (0%)
Live where						
Home	315 (99.1%)	136 (98.6%)	41 (97.6%)	11 (100%)	51 (100%)	76 (100%)
Residence for eldery (RPA)	2 (0.6%)	2 (1.4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Nursing home (CHSLD)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Intermediate and family-type ressources (RI-RTF)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
In rooming house	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Missing n(%)	1 (0.3%)	0 (0%)	1 (2.4%)	0 (0%)	0 (0%)	0 (0%)
Live with						
Family member(s)	279 (87.7%)	123 (89.1%)	38 (90.5%)	11 (100%)	45 (88.2%)	62 (81.6%)
Caretaker	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Alone	31 (9.7%)	10 (7.2%)	4 (9.5%)	0 (0%)	5 (9.8%)	12 (15.8%)
Roommate(s)	4 (1.3%)	2 (1.4%)	0 (0%)	0 (0%)	1 (2.0%)	1 (1.3%)
Inconnu n(%)	4 (1.3%)	3 (2.2%)	0 (0%)	0 (0%)	0 (0%)	1 (1.3%)
COVID severity						
Mild	299 (94.0%)	123 (89.1%)	40 (95.2%)	11 (100%)	51 (100%)	74 (97.4%)
Moderate	11 (3.5%)	11 (8.0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Severe	8 (2.5%)	4 (2.9%)	2 (4.8%)	0 (0%)	0 (0%)	2 (2.6%)
Death	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Infection's wave						
1	22 (6.9%)	16 (11.6%)	6 (14.3%)	0 (0%)	0 (0%)	0 (0%)
2	130 (40.9%)	110 (79.7%)	15 (35.7%)	5 (45.5%)	0 (0%)	0 (0%)
3	22 (6.9%)	11 (8.0%)	7 (16.7%)	4 (36.4%)	0 (0%)	0 (0%)
4	26 (8.2%)	1 (0.7%)	7 (16.7%)	1 (9.1%)	16 (31.4%)	1 (1.3%)
5	56 (17.6%)	0 (0%)	3 (7.1%)	1 (9.1%)	30 (58.8%)	22 (28.9%)
6	16 (5.0%)	0 (0%)	1 (2.4%)	0 (0%)	2 (3.9%)	13 (17.1%)
7	46 (14.5%)	0 (0%)	3 (7.1%)	0 (0%)	3 (5.9%)	40 (52.6%)
Reinfection's wave						
1	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
2	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
3	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
4	1 (0.3%)	1 (0.7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
5	72 (22.6%)	45 (32.6%)	19 (45.2%)	5 (45.5%)	3 (5.9%)	0 (0%)

Characteristics	Cohort (n = 318)	Clusters				
		1 (n = 138)	2 (n = 42)	3 (n = 11)	4 (n = 51)	5 n = 76)
6	43 (13.5%)	32 (23.2%)	7 (16.7%)	0 (0%)	4 (7.8%)	0 (0%)
7	202 (63.5%)	60 (43.5%)	16 (38.1%)	6 (54.5%)	44 (86.3%)	76 (100%)
Delay between events (days): mean (SD)						
Vaccine1-Vaccine2	103 (63.6)	110 (70.5)	387 (43.8)	129 (84.5)	78.5 (24.3)	90.3 (23.9)
Vaccine1-Vaccine3	298 (67.8)	290 (76.4)	--	331 (66.6)	331 (77.4)	293 (49.0)
Vaccine1-Infection	70.7 (267)	-151 (91.9)	-173 (118)	49.5 (70.8)	255 (81.0)	416 (83.5)
Vaccine1-Reinfection	463 (172)	379 (114)	247 (141)	405 (108)	518 (123)	648 (112)
Vaccine1-Reinfection2	523 (151)	494 (90.9)	413 (114)	--	--	747 (177)
Vaccine1-Reinfection3	583 (113)	553 (133)	--	--	--	643 (21.9)
Vaccine2-Vaccine3	202 (60.1)	189 (67.5)	--	199 (65.1)	260 (71.6)	202 (32.0)
Vaccine2-Infection	-18.6 (285)	-261 (103)	-666 (163)	-79.1 (34.2)	177 (72.5)	326 (72.3)
Vaccine2-Reinfection	373 (182)	268 (121)	-213 (125)	277 (124)	439 (121)	557 (102)
Vaccine2-Reinfection2	407 (190)	371 (107)	-100 (--)	--	--	656 (150)
Vaccine2-Reinfection3	467 (101)	410 (62.7)	--	--	--	581 (18.4)
Vaccine3-Infection	-169 (269)	-427 (103)	--	-278 (74.4)	-96.0 (45.8)	125 (62.7)
Vaccine3-Reinfection	213 (166)	111 (134)	--	88.7 (129)	179 (129)	357 (93.6)
Vaccine3-Reinfection2	245 (169)	171 (98.0)	--	--	--	486 (111)
Vaccine3-Reinfection3	315 (92.1)	275 (84.9)	--	--	--	396 (NA)
Infection-Reinfection	391 (177)	529 (118)	392 (183)	356(126)	263 (103)	232 (84.2)
Infection-Reinfection2	564 (154)	614 (107)	593 (178)	--	--	330 (48.0)
Infection-Reinfection3	583 (170)	686 (75.2)	--	--	--	376 (7.07)
Reinfection-Reinfection2	175 (86.7)	162 (88.5)	250 (59.3)	--	--	139 (61.4)
Reinfection-Reinfection3	210 (69.6)	222 (78.4)	--	--	--	186 (62.9)
Reinfection2-Reinfection3	96.5 (69.6)	102 (88.4)	--	--	--	85.5 (19.1)
Doses of vaccine at primary infection: mean (SD)	1.08 (1.30)	0 (0)	0.024 (0.154)	1 (0)	2.02 (0.140)	2.99 (0.115)
Doses of vaccine at first reinfection: mean (SD)	2.36 (0.876)	2.55 (0.514)	0.50 (0.506)	2.64(0.505)	2.39 (0.493)	2.99 (0.115)
Doses of vaccine at second reinfection: mean (SD)	2.29 (0.864)	2.60 (0.503)	0.83 (0.408)	--	--	2.80 (0.447)
Doses of vaccine at third reinfection: mean (SD)	2.50 (0.548)	2.50 (0.577)	--	--	--	2.50 (0.707)
Number of reinfections: mean (SD)	1.12 (0.375)	1.17 (0.451)	1.14 (0.354)	1.00 (0)	1.00 (0)	1.09 (0.372)

The Figure 1 presents the global sequence of events identified for the cohort. Boxes represent the events while the edges represent the temporal sequence between the events. All boxes and edges are completed with the relative frequency of each event or transition as well as the median time between each transition. We used the sum of the medians of the various transitions as an approximation of sequence duration for illustrative purposes. While the medians are not addable due to their statistical properties and do not represent the actual median of the complete sequence, they can support the identification of overall trends. To avoid confusion, this approximation method will hereafter be referred to as *summed medians*. The mapping shows that 56.3% of individuals in the cohort began their sequence with the primary infection, while 43.7% started with

a first dose of vaccine. Of all those who received the first dose, regardless of the previous event, 83.3% were subsequently vaccinated a second time, within a median of 87 days. Among those who received a second dose, regardless of previous trajectory, 50.3% then received a third vaccine within a median of 191 days, while 16% contracted their first infection within a median of 183 days.

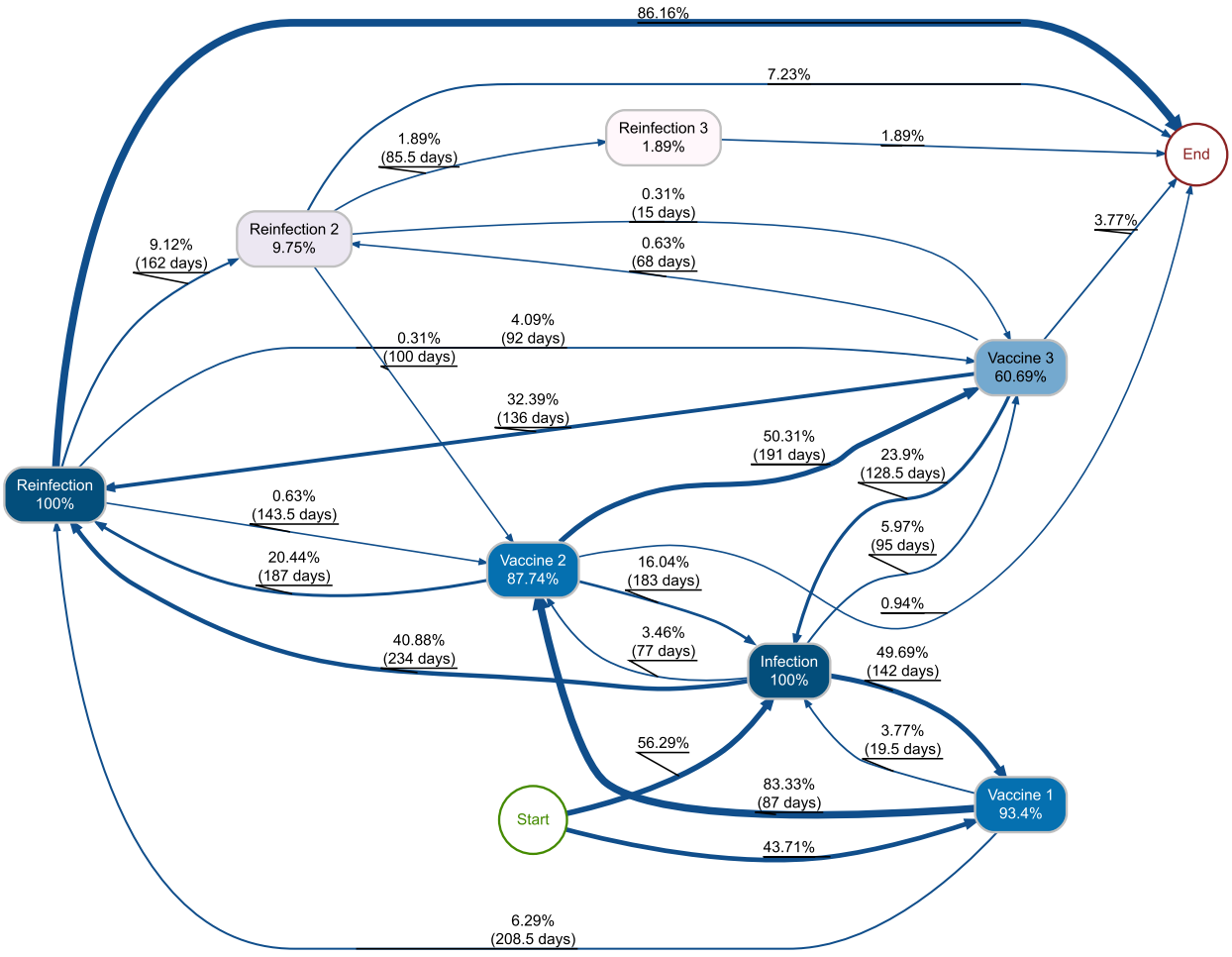


Figure 1 – Sequencing map of infection, reinfection and vaccination events in the cohort. The map has been generated using process mining techniques. Boxes indicate main events and their relative frequency through the whole cohort. Edges represent the consecutive sequence of events and display the median time between each event as well as the relative frequence of participants following a specific sequence of events.

Socio-demographic, participant's condition, habits and environment characteristics

To achieve the objective of grouping participants by infection, vaccination, and reinfection patterns, we used event sequences as time series in a cluster analysis, resulting in five distinct groups. The first cluster included 138 participants (43,4%), while others respectively included 42 (13,2%), 11 (3,5%), 51 (16%) and 76 individuals (23,9%). There was no significant statistical difference between clusters in terms of age (p=0,137) and body mass index (BMI) (p=0,545), and the proportion of males and females in each cluster was similar to the cohort proportion. However, in the first three clusters, the ratio of health workers was lower (between 21,4% to 36,4%) than the cohort proportion (44,3%), while in the last two clusters, the ratio was reversed, with 51% and 76.3%. The detailed characteristics of each group are shown in Table 1.

Temporal description

The following section will describe each group in terms of temporal sequence and present the illustrations of these sequences (Figure 2A to Figure 2E).

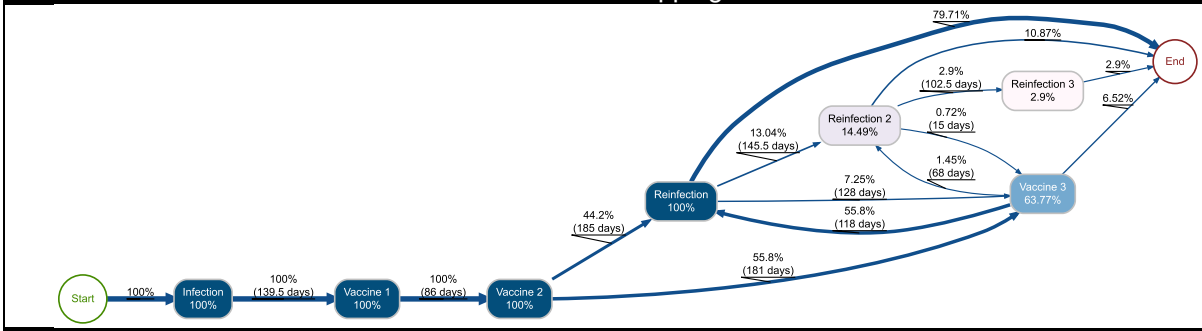
Cluster 1. Patients in the first cluster were mostly infected in the first three waves of the pandemic (Table 1). As shown in Figure 2A, which presents the event mapping for this group, individuals were first infected before receiving two doses of vaccine within a *summed medians* of 225.5 days (median I-V1 139.5; median V1-V2 86). The sequence then split, with 44.2% of the group who were reinfected, while the remaining received their third dose of vaccine, both events within similar median timescales (185 and 181 days respectively). Patients who received this last dose of vaccine took a median 118 days before being reinfected. Reinfection occurred in waves five, six and seven of the pandemic.

Cluster 2. In common with group 1, most of individuals in this group were infected in the first waves (66.7%), although a few individuals contracted their primary infection later in the pandemic (Table 1). Almost the entire group (97.6%) was infected before a first dose of vaccine (Figure 2B). The sequence of the entire group converged towards a reinfection, either following first dose (47.6%, *summed medians* 380.5 days (172; 208.8) from primary infection), or following the initial

infection (52.4%, median 270 days). Some individuals had contracted the disease two times without vaccine in their sequences, which means that this group has the particularity to contain patient with natural immunity instead of hybrid immunity (N=21). Reinfection occurred in waves five, six and seven, but mainly in five and six (61,9%).

A

Events mapping

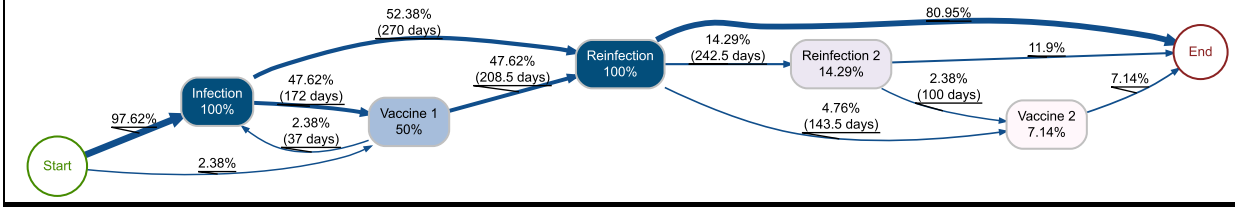


Events sequences

1	2	3	4	5	6	7	n	Cumulative frequency (%)
I	V1	V2	V3	R1			67	48,55
I	V1	V2	R1				43	79,71
I	V1	V2	R1	V3			8	85,51
I	V1	V2	V3	R1	R2		8	91,31
I	V1	V2	R1	R2			5	94,93
I	V1	V2	R1	R2	R3		2	96,38
I	V1	V2	V3	R1	R2	R3	2	97,83
I	V1	V2	R1	V3	R2		2	99,28
I	V1	V2	R1	R2	V3		1	100,00

B

Events mapping

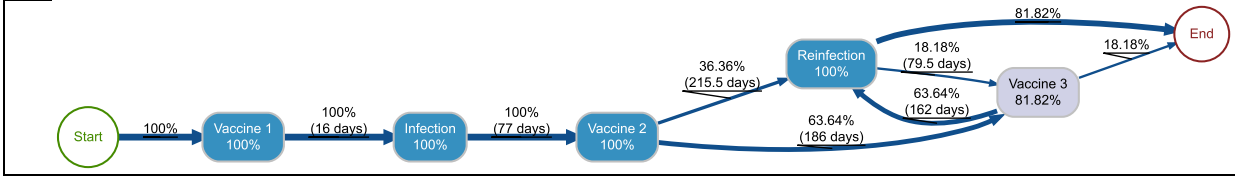


Events sequences

1	2	3	4	5	6	7	n	Cumulative frequency (%)
I	R1						20	47,62
I	V1	R1					13	78,57
I	V1	R1	R2				4	88,10
I	V1	R1	V2				2	92,86
I	V1	R1	R2	V2			1	95,24
V1	I	R1					1	97,62
I	R1	R2					1	100,00

C

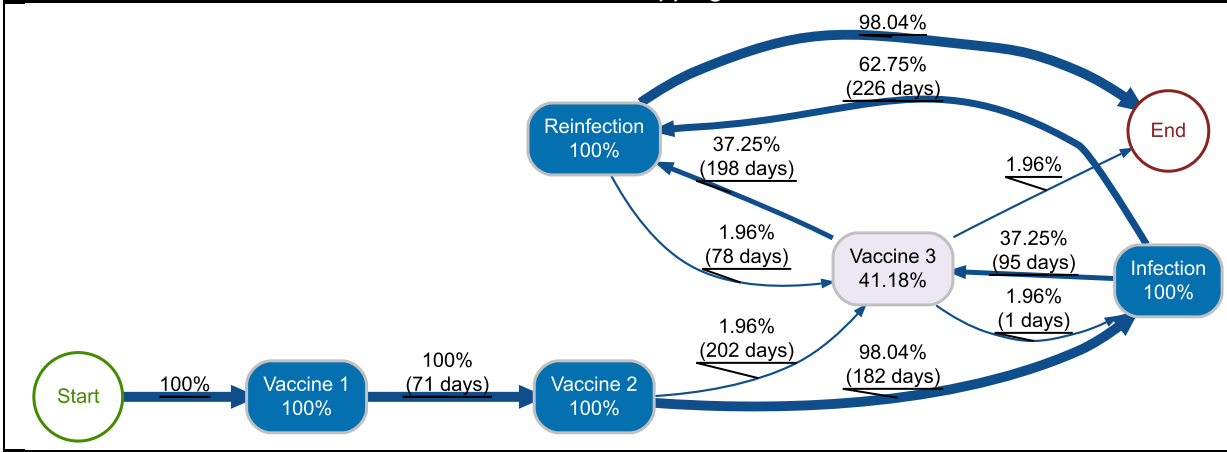
Events mapping



Events sequences								
1	2	3	4	5	6	7	n	Cumulative frequency (%)
V1	I	V2	V3	R1			7	63,64
V1	I	V2	R1				2	81,82
V1	I	V2	R1	V3			2	100,00

D

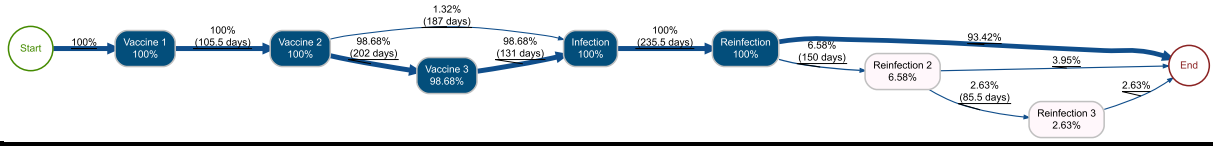
Events mapping



Events sequences								
1	2	3	4	5	6	7	n	Cumulative frequency (%)
V1	V2	I	R1				30	58,82
V1	V2	I	V3	R1			19	96,08
V1	V2	I	R1	V3			1	98,04
V1	V2	V3	I	R1			1	100,00

E

Events mapping



Events sequences								
1	2	3	4	5	6	7	n	Cumulative frequency (%)
V1	V2	V3	I	R1			71	93,42
V1	V2	V3	I	R1	R2		3	97,37
V1	V2	V3	I	R1	R2	R3	1	98,68
V1	V2	I	R1	R2	R3		1	100,00

Legend :
I : Infection
V1 : 1st dose of vaccine
V2 : 2nd dose of vaccine
V3 : 3nd dose of vaccine
R1 : 1st reinfection
R2 : 2nd reinfection
R3 : 3nd reinfection

Figure 2 – Sequences and mapping of interest events, grouped by cluster where A) is cluster 1, B) cluster 2, C) cluster 3, D) cluster 4 and E) cluster 5. Events mappings present the relative frequency of the trajectory (%) and the median in days between each event.

Cluster 3. In this group, individuals were mainly infected in the second and third waves of the pandemic (81.9%), none having been infected in the first wave (Table 1). Thus, on average, they were infected slightly later than the previous two groups. This is interesting given that, as Figure 2C shows, the entire group started their event sequence with the first vaccine. However, the delay between this vaccine and primary infection is relatively short, with a mean and median delay of 49.5 and 16 days. Post-infection, individuals in the group all received a second vaccine before they split into two subsequences: those who were reinfected (36.4%) within a median of 215.5 days and those who received a third dose of vaccine (63.6%) within a median of 186 days. Those individuals took a median delay of 162 additional days before their reinfection. Within this group, 45.5% of individuals were reinfected in the fifth wave, the remaining 54.5% were reinfected after the sixth wave, meaning that no reinfection occurred in wave 6 for this group.

Cluster 4. The fourth group was mainly infected in the fourth (31.4%) and the fifth wave (58.8%), positioning the group, in terms of timeline of infection, between cluster 3 and cluster 5 (Table 1). For the entire group, the sequence of events begins with two vaccines, as shown in Figure 2D. For a majority (98%), the subsequent event is the primary infection. Once this event is reached, the group separates into two distinct trajectories: towards the third vaccine dose (37.3%, median delay 95 days) or the reinfection (62.8%, median delay 226 days). Individuals who received the 3rd dose after their primary infection were reinfected within a median of 198 days. In terms of median delay from first vaccination, individuals with the V1-V2-I-R sequence had a *summed medians* delay of 479 days (71; 182; 226 days), compared with patients with the V1-V2-I-V3-R sequence, who had a *summed medians* delay of 546 days (71; 182; 95 198 days). Reinfections for this group occurred mainly from the seventh wave onwards. There is in the group an individual with the same sequence of majority of the group 5 (V1-V2-V3-I-R1). Analysis showed that this individual, even if it has the same sequence, got his third vaccine only one day before is first infection, making it more like the majority of group 4 (V1-V2-I-R1).

Cluster 5. The last group contains the individuals who were most vaccinated prior to their primary infection. Indeed, 98.7% had received their third dose at the time of initial infection. Reinfection

occurred latest among the other four groups, i.e. during waves five, six and mainly from wave seven onwards (52.6%). This represents a median delay from infection of 235.5 days.

In addition, we noted interesting statistical differences between the groups. As presented before, at the time of their primary infection, individuals in the first cluster did not received a vaccine, individuals in the third had mostly received one dose, individuals in the fourth, two doses (average 2.02; 95% confidence interval 1.981-2.059) and in the fifth, three doses (2.99; 2.964-3.016). This situation presented a strong statistical difference between groups ($p<0.001$), except for clusters one with two ($p=8.76e-1$) and three with four ($p=2.32e-1$). About doses at first reinfection, there was evidence of strong statistical difference of average doses of vaccine, excluding for cluster 1 (2.55; 2,464-2,636) with 3 (2.64; 2.301-2.979) and 4 (2.39; 2.251-2.529), and cluster 3 with cluster 4 and 5 (2.99; 2.964-3.016). In clusters containing individuals who were reinfected two times (clusters 1, 2 and 5), there was a significant statistical difference ($p<0.001$) between doses at second reinfection for clusters 1 and 2 and between clusters 2 and 5 ($p<0.001$). However, for individuals who were reinfected three times (clusters 1 and 5), there was no evidence of statistical difference between groups ($p=1$).

Follow up duration

In terms of length of follow-up, an analysis of the distributions revealed certain disparities. Group 1 showed a moderate distribution around the median (729 days), but was distinguished by the presence of outliers, suggesting that some participants in this group had below- and above-average follow-up times. That group had a statistical difference with group 2-4-5 (p-value < 0.001). Group 3 stood out for its good homogeneity, low variability and absence of outliers, suggesting uniform follow-up (interquartile range (IQR) = 190, median 694 days) and had no statistical difference with any of the other groups. Group 2, on the other hand, shows the largest heterogeneity, with a much wider interquartile range (479 days), suggesting significant differences in follow-up duration between participants in this group. Despite showing similar visual distribution with close medians and moderate variability (median of 562 and 676 days; IQR of 191 and 151 days), there was a significant difference in follow-up duration between groups 4 and 5 (p-value<0.001).

In summary, the cohort participants were grouped into five clusters, and their characterization revealed that the clusters follow a temporal progression according to the timing of infection and its positioning across the waves of the pandemic. Reinfections, on the other hand, occurred from the fifth wave onwards. The most highly vaccinated groups appear to have been infected, and consequently reinfected, later in the pandemic. Some groups featured a higher proportion of healthcare workers, while for others, it was the trajectory and their timeframes that were of interest. There were some disparities in follow-up times, which will need to be taken into account when drawing conclusions from the results.

1.4 Discussion

The project aimed to study hybrid immunity by identifying and characterizing SARS-CoV-2 reinfection profiles. Using machine learning techniques, we grouped individuals from the BQC19 according to characteristics leading to similar pattern of vaccination, infection, and reinfection in a five-cluster classification.

The study showed no significant differences between the groups regarding socio-demographic variables, except for the proportion of healthcare workers. For this variable, groups 4 (51%) and 5 (76.3%) had a higher proportion than the cohort (44.3%). These same groups had a more sustained initial vaccination sequence than the other groups (2 doses and 3 doses, respectively before primary infection). This seemed consistent with the vaccination policies in place during the pandemic for this at-risk population in close contact with the virus. The results therefore suggest that these policies had a positive impact, given that, for this group, primary infection occurred later during the pandemic. However, this finding, implying that healthcare workers in the cohort were infected late (59.6% of them), differs from the results of Carazo et al. (2023) in their study about healthcare workers' protection against Omicron BA.2 reinfection conferred depending on the primary infection variant, where, for around the same period, approximately 20.7% of healthcare workers were infected. The difference may be explained by the inclusion criteria for the documented dates in our study, which reduced the size of our cohort. This is in contrast to their study, which exploited data sources from the Ministry of Health and Social Services that were potentially more exhaustive at this level.

Similarly, it was possible to observe that group 3, which received a first vaccine before being infected, was spared in wave 1. Thus, compared with groups 1 and 2, who received no vaccine prior to infection, group 3's primary infection occurred later in waves 2 and 3, allowing us to hypothesize that, although one dose was missing to achieve so-called complete vaccine immunization, the first vaccine dose may have generated a positive impact by delaying the initial infection. However, this hypothesis must be interpreted with caution, given the small number of individuals in the group.

The first group also presented interesting features in terms of vaccination efficacy. Indeed, following the second vaccine, the trajectory of the individuals split in two, some towards reinfection (median of 185 days after) while the others towards the third vaccine (median of 181 days after). This separation occurred within an almost identical median time, which might suggest that the policy of administering the third dose was relatively synchronized with a weakening of immunity. This timeframe is in line with the results of Asamoah-Boaheng et al. (2023) showing that antibody levels decrease with a half-life of 94 days and plateauing at 294 days. Although these results relate to mRNA vaccines, and vaccine type was not a variable in the present project, the results remain consistent. Also, the 3rd vaccine also appears to have delayed reinfection by 118 days (median) compared with patients who received only 2 doses. This suggests that an earlier 3rd dose could potentially have prevented more reinfections. Group 3 had a similar separation between the reinfection event and the third vaccine. The latter, whose sequence prior to separation was V1-I-V2, compared with the first group's I-V1-V2, had a greater *summed medians* time to reinfection (215 vs. 185 days). Similarly, when comparing the median time from second vaccine to reinfection via third vaccine dose, group 3 had a longer *summed medians* time (348 vs. 299 days). This might suggest that, in terms of hybrid immunity, being infected between two vaccine doses could offer slightly longer-lasting immunity, but the size of the 3rd group makes it difficult to draw such a definitive conclusion.

The study also revealed a group of patients who had not been vaccinated and, consequently, had not achieved hybrid immunity. This same group also contained individuals who had received only one vaccine, implying that they had not fully achieved hybrid immunity, given that full vaccine immunity required 2 doses. Thus, based exclusively on the sequence of events and their temporality, individuals with partial or non-existent hybrid immunity were grouped together by

the clustering algorithm. Despite its interest, this finding needed to be nuanced according to the variance of follow-up duration within the group. In fact, some individuals may simply not have had the follow-up time required to be fully vaccinated. Despite this limitation, it is worth mentioning that the algorithm's data-driven grouping of these participants supports the finding of Sanchez-de Prada et al. (2024) that there is no significant difference between individuals vaccinated once within five months of infection and those who were not vaccinated at all.

The results showed that vaccination has a positive effect in delaying infection or reinfection. They also showed that the temporality of events greatly influenced the formation of groups by the algorithm, in the sense that primary infections and reinfections are distributed according to a temporal progression, from group 1 (the earliest infections) to group 5 (the latest).

In terms of strengths, we used data collected as early as the beginning of the pandemic, which allowed us to use valuable data for this study. The data management process is also a great strength for this study as substantial work has been performed in order to consolidate the data, allowing us to increase our sample size. Finally, the use of machine learning made it possible to identify more complex patterns by taking events and their temporality into account, using a data-driven approach. In fact, the method enabled us to consider not only the delay between events, but also their chronology, in order to take into account the pandemic’s waves. Thus, by first forming the groups, then characterizing them using variables that were not used in the clustering process, we were able to highlight elements that were more difficult to identify using conventional methods. Non-supervised techniques reveal interesting avenues for future investigation. To do so, we intend to use genomic data to characterize the groups, including the use of random forest to determine the most relevant variables for this purpose.

While there are strengths, there are also some limitations to this work. First, even though the significant attention given to data management, the sample size remained small. This is mostly due to our inclusion criteria, where people need to be reinfected to be included. Second, while the infection and vaccinations dates were properly collected within the datasets, we had to establish a strategy to correct the reinfection date as two variables were presents within the same dataset. However, dates were the same for the vast majority of participants, and the same treatment was applied otherwise, limiting potential biases. In addition, the delay used to define the reinfection in

this study differed from the delay found in the literature and could be considered as a limitation. Considering the lack of official consensus in the scientific community regarding this timeframe, the choice of timeframe (14 days) has been made using the threshold used among the BQC19 research community to enable meaningful comparisons. However, even with this short delay, the number of patients concerned was relatively small and, according to the distribution of delay between reinfections, increasing the threshold of consideration would have only a minimal impact on the number of individuals. Finally, the use of *summed medians* can induce a distortion in the estimation of overall times. However, it was only used to give an overall idea of the temporality of the sequences and the global trajectory, as we know that is not the actual median of the sequence.

1.4 Conclusion

In conclusion, the study highlights the role of vaccination, in line with current knowledge. It also shows that, beyond the sequence of events, it is rather their temporality and the delays between them that are of greatest importance. In terms of hybrid immunity, the study suggests that an infection between two vaccines could offer greater immunity. This finding should be treated with caution, however, given the size of the group from which it is drawn and the disparity in follow-up times. In any case, this is an interesting perspective to pursue.

1.5 Contributions to knowledge

What does this study add to existing knowledge?

To our knowledge, this is the first study using data from the Biobanque québécoise de la COVID-19 to investigate reinfection patterns and hybrid immunity using a data-driven approach. In addition to highlighting the effectiveness of vaccination policies, it identified, by leveraging machine learning technics on complex multidimensional timeseries, distinct groups and COVID-19 patterns of infection, reinfection and vaccination, thus providing interesting insights for further investigation. It also highlights that beyond the sequence of events, the temporal delays between events seem to play an important role in the acquisition of primary and secondary infections.

What are the key implications for public health interventions, practice or policy?

Delays between events played a determining role in the formation of the study groups. Consequently, their consideration in the development and adaptation of health policies, particularly regarding vaccine administration and boosters, is necessary. In addition, the study shows that machine learning algorithms represent, for public health practices, an innovative and complementary approach to analyze health data and discover hidden information that can have impact on public health decisions. These two avenues, combining delay analysis and machine learning, offer promising perspectives for future work, particularly in preparation for possible pandemics.

References

Asamoah-Boaheng, M., B. Grunau, S. Haig, M. E. Karim, T. Kirkham, P. M. Lavoie, et al. (2023). « Eleven-month SARS-CoV-2 binding antibody decay, and associated factors, among mRNA vaccinees: implications for booster vaccination », Access Microbiol, vol. 5, no 11.

Carazo, S., D. M. Skowronski, M. Brisson, S. Barkati, C. Sauvageau, N. Brousseau, et al. (2023). « Protection against omicron (B.1.1.529) BA.2 reinfection conferred by primary omicron BA.1 or pre-omicron SARS-CoV-2 infection among health-care workers with and without mRNA vaccination: a test-negative case-control study », Lancet Infect Dis, vol. 23, no 1, p. 45-55.

Giorgino, T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. Journal of Statistical Software, 31(7), 1 - 24. https://doi.org/10.18637/jss.v031.i07

Hu, B., Guo, H., Zhou, P., & Shi, Z. L. (2021). Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol, 19(3), 141-154. https://doi.org/10.1038/s41579-020-00459-7

Janssenswillen, G., Depaire, B., Swennen, M., Jans, M., & Vanhoof, K. (2019). bupaR: Enabling reproducible business process analysis. Knowledge-Based Systems, 163, 927-930. https://doi.org/https://doi.org/10.1016/j.knosys.2018.10.018

Lapuente, D., Winkler, T. H., & Tenbusch, M. (2024). B-cell and antibody responses to SARS-CoV-2: infection, vaccination, and hybrid immunity. Cell Mol Immunol, 21(2), 144-158. https://doi.org/10.1038/s41423-023-01095-w

Livieratos, A., Gogos, C., & Akinosoglou, K. (2024). Impact of Prior COVID-19 Immunization and/or Prior Infection on Immune Responses and Clinical Outcomes. Viruses, 16(5). https://doi.org/10.3390/v16050685

Meyer, D., & Buchta, C. (2022). proxy: Distance and Similarity Measures. R Package. In (Version 0.4-27) https://cran.r-project.org/web/packages/proxy/index.html

Misra, A., & Theel, E. S. (2022). Immunity to SARS-CoV-2: What Do We Know and Should We Be Testing for It? J Clin Microbiol, 60(6), e0048221. https://doi.org/10.1128/jcm.00482-21

Posit team. (2024). RStudio: Integrated Development Environment for R. In Posit Software. http://www.posit.co/.

R Core Team. (2023). R : A Language and Environment for Statistical Computing. In R Foundation for Statistical Computing. https://www.R-project.org/

Rodriguez Velásquez, S., Biru, L. E., Hakiza, S. M., Al-Gobari, M., Triulzi, I., Dalal, J., Varela, C. B. G., Botero Mesa, S., & Keiser, O. (2024). Long-term levels of protection of different types of immunity against the Omicron variant: a rapid literature review. Swiss Med Wkly, 154, 3732. https://doi.org/10.57187/s.3732

Rousseeuw, Peter J. (1987). « Silhouettes: A graphical aid to the interpretation and validation of cluster analysis », Journal of Computational and Applied Mathematics, vol. 20, p. 53-65.

Sakoe, H. et S. Chiba (1978). « Dynamic programming algorithm optimization for spoken word recognition », IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no 1, p. 43-49.

Sarda-Espinosa, A. (2023). dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance. R package In (Version 5.5.12) https://CRAN.R-project.org/package=dtwclust

Tremblay, K., Rousseau, S., Zawati, M. n. H., Auld, D., Chassé, M., Coderre, D., Falcone, E. L., Gauthier, N., Grandvaux, N., Gros-Louis, F., Jabet, C., Joly, Y., Kaufmann, D. E., Laprise, C., Larochelle, C., Maltais, F., Mes-Masson, A.-M., Montpetit, A., Piché, A. on behalf of BQC19. (2021). The Biobanque québécoise de la COVID-19 (BQC19)—A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories. PLOS ONE, 16(5), e0245031. https://doi.org/10.1371/journal.pone.0245031

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Yutani, H. (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43). https://doi.org/10.21105/joss.01686

World Health Organization. (2024). WHO COVID-19 dashboard - Data reported on 28 april 2024. Retrieved 17 juillet 2024 from https://data.who.int/dashboards/covid19/

World Health Organization, & Kryuchkov, I. (2022, 7 December). Post COVID-19 condition (Long COVID). Retrieved 2024, September 9 from https://www.who.int/europe/news-room/fact-sheets/item/post-covid-19-condition

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., Tan, W., China Novel Coronavirus, I., & Research, T. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med, 382(8), 727-733. https://doi.org/10.1056/NEJMoa2001017

Chapitre 5 Résultats supplémentaires

Les résultats de la classification hiérarchique ascendante sur une matrice de distance obtenue avec l'algorithme DTW (modèle HA-DTW) ayant été présentés au Chapitre 4, le présent chapitre portera sur les résultats des autres analyses de regroupement effectuées en fonction des séquences temporelles, mais non présentées dans l'article. Ces derniers ont été obtenus sur la même cohorte que celle présentée au chapitre précédent dont le diagramme de flux d'inclusion est présenté en Annexe 5.

Afin d'avoir une représentativité des algorithmes et des différentes mesures de dissimilarité de l'étude, trois modèles ont été retenus sur la base du coefficient de silhouette moyen (sil), à savoir : 1) une classification hiérarchique **descendante** sur une matrice de distance **Euclidienne** calculée sur les délais interévénements **non normalisés** (modèle HD, sil=0,58), 2) une classification hiérarchique **ascendante** sur une matrice de distance de **Manhattan** sur les délais interévénements **normalisés** (modèle HA-Norm, sil=0,50) et 3) une classification hiérarchique **descendante** sur une matrice de distance obtenue sur les données des **événements** à partir de l'algorithme ***Optimal matching*** (HD-OM, sil=0,12). Les trois algorithmes ont classé les patients en deux groupes, respectivement de 176 et 142, 178 et 140 puis 291 et 27 individus.

5.1 Séquences d'événements

Pour rappel, une séquence est l'enchaînement des événements d'intérêt (l'infection, la ou les vaccinations, la ou les réinfections) de chaque individu. Ainsi, les séquences des événements sont particulièrement utiles à l'atteinte de l'objectif du projet d'identifier et caractériser les différents profils de réinfection. À cet effet, les différentes séquences d'événements sont présentées dans le Tableau 4.

5.1.1 Modèles HD et HA-Norm

En ce qui a trait aux plus fréquentes, soit celles représentant une fréquence cumulée de 75%, les séquences étaient similaires pour les modèles HD et HA-Norm. Ainsi, pour leur groupe 1 (respectivement Figure 5A et Figure 5C), la séquence la plus fréquente était

composée successivement de l'infection primaire (I), des trois doses de vaccins (V1-V2-V3), puis de la première réinfection (R1).

Tableau 4 - Séquence des événements des patients de la cohorte selon les groupes des modèles HD, HA-Norm et HD-OM

Séquence des événements : n (%)							Cohorte (n = 318)	HD		HA-Norm		HD-OM	
								Groupe 1 (n = 176)	Groupe 2 (n = 142)	Groupe 1 (n = 178)	Groupe 2 (n = 140)	Groupe 1 (n = 291)	Groupe 2 (n = 27)
V1	V2	V3	I	R1			72 (22,64)	--	72 (50,7)	--	72 (51,5)	71 (24,4)	1 (3,7)
I	V1	V2	V3	R1			67 (21,07)	66 (37,5)	1 (0,7)	67 (37,6)	--	57 (19,6)	10 (37,0)
I	V1	V2	R1				43 (13,52)	43 (24,4)	--	43 (24,3)	--	43 (14,8)	--
V1	V2	I	R1				30 (9,43)	--	30 (21,1)	--	30 (21,5)	30 (10,3)	--
I	R						20 (9,29)	10 (5,7)	10 (7,0)	9 (5,2)	11 (7,9)	20 (6,9)	--
V1	V2	I	V3	R1			19 (5,97)	--	19 (13,5)	--	19 (13,7)	15 (5,2)	4 (14,8)
I	V1	R1					13 (4,09)	13 (7,4)	--	13 (7,3)	--	13 (4,5)	--
I	V1	V2	R1	V3			8 (2,52)	8 (4,6)	--	8 (4,5)	--	8 (2,8)	--
I	V1	V2	V3	R1	R2		8 (2,52)	8 (4,6)	--	8 (4,5)	--	5 (1,7)	3 (11,1)
V1	I	V2	V3	R1			7 (2,20)	4 (2,3)	3 (2,1)	6 (3,4)		5 (1,7)	2 (7,4)
I	V1	V2	R1	R2			5 (1,57)	5 (2,8)	--	5 (2,8)	--	5 (1,7)	--
I	V1	R1	R2				4 (1,26)	4 (2,3)	--	4 (2,3)	--	4 (1,4)	--
V1	V2	V3	I	R1	R2		3 (0,94)	--	3 (2,1)	--	3 (2,2)	3 (1,0)	--
I	V1	R1	V2				2 (0,64)	2 (1,1)	--	2 (1,1)	--	2 (0,7)	--
I	V1	V2	R1	R2	R3		2 (0,64)	2 (1,1)	--	2 (1,1)	--	2 (0,7)	--
I	V1	V2	V3	R1	R2	R3	2 (0,64)	2 (1,1)	--	2 (1,1)	--	--	2 (7,4)
V1	I	V2	R1				2 (0,64)	2 (1,1)	--	2 (1,1)	--	2 (0,7)	--
V1	I	V2	R1	V3			2 (0,64)	2 (1,1)	--	2 (1,1)	--	2 (0,7)	1 (3,7)
I	V1	V2	R1	V3	R2		2 (0,64)	2 (1,1)	--	2 (1,1)	--	1 (0,3)	1 (3,7)
I	V1	V2	R1	R2	V3		1 (0,31)	1 (0,6)	--	1 (0,5)	--	--	1 (3,7)
I	V1	R1	R2	V2			1 (0,31)	1 (0,6)	--	1 (0,5)	--	1 (0,3)	--
V1	V2	I	R1	V3			1 (0,31)	--	1 (0,7)	--	1 (0,8)	1 (0,3)	--
V1	V2	V3	I	R1	R2	R3	1 (0,31)	--	1 (0,7)	--	1 (0,8)	--	1 (3,7)
V1	V2	I	R1	R2	R3		1 (0,31)	--	1 (0,7)	--	1 (0,8)	--	1 (3,7)
V1	I	R1					1 (0,31)	1 (0,6)	--	1 (0,5)	--	1 (0,3)	--
I	R1	R2					1 (0,31)	--	1 (0,7)	--	1 (0,8)	1 (0,3)	--

Elle représente, dans l'ordre, 66 et 67 individus, soit 37,5% et 37,6% des séquences de leur groupe. Parmi les autres séquences fréquentes, un nombre de trois séquences supplémentaires composaient le groupe 1 des modèles à savoir I-V1-V2-R1 (n=42), I-V1-R1 (n=13) et finalement I-R1 (n=10 pour le modèle HD et n=9 pour HA-Norm). À noter que la séquence IR-1 suggérait que ces individus, n'ayant pas de vaccins dans leur séquence, avaient une immunité naturelle plutôt qu'hybride. Les différents profils de réinfection les plus fréquents des groupes 1 des deux modèles basés sur les délais interévénements démarraient tous leur séquence avec l'infection primaire.

Pour les groupes 2 de ces mêmes modèles, trois séquences fréquentes ont été identifiés, à savoir V1-V2-V3-I-R1 (n=72), V1-V2-I-R1 (n=30) et V1-V2-I-V3-R1 (n=19). Pour les deux modèles, les séquences les plus fréquentes avaient toutes comme point de départ l'enchaînement des deux premières doses de vaccin. La Figure 5B présente, pour le groupe 2, les séquences du modèle HD alors que la Figure 5D, celles du modèle HA-Norm.

5.1.2 Modèle HD-OM

Relativement au modèle HD-OM, ces deux groupes présentaient chacun cinq séquences fréquentes. Ces derniers sont illustrés à la Figure 6. Pour le premier groupe, 71 individus avaient la séquence V1-V2-V3-I-R1, soit la totalité des individus de la cohorte ayant cette séquence. Également, 57 patients avec la séquence I-V1-V2-V3-R1 ont été classés dans le groupe 1 bien que les 10 autres avec la même séquence se retrouvaient dans le deuxième groupe. Dans ce deuxième groupe, il n'y avait qu'un seul travailleur de la santé avec cette séquence, les 23 autres étant tous dans le groupe 1. Puisque le statut de travailleur de la santé n'expliquait pas la raison de leur séparation alors que la séquence était identique, nous avons analysé les délais entre les événements. La seule distinction statistiquement significative (p=0,01) entre les deux groupes était au niveau du délai entre la 2^{e} dose de vaccin et l'infection primaire qui semblait avoir été en moyenne plus court pour le second groupe avec 194 jours (IC 95% : 163,1-224,9) versus 244 jours pour le premier (IC 95% : 219,4,1-268,6).

A *Modèle HD*

Événements							n	Fréquence cumulée (%)
1	2	3	4	5	6	7		
I	V1	V2	V3	R1			66	37,50
I	V1	V2	R1				43	61,93
I	V1	R1					13	69,32
I	R1						10	75,00

B *Modèle HD*

V1	V2	V3	I	R1			72	51,70
V1	V2	I	R1				30	71,83
V1	V2	I	V3	R1			19	85,21

C *Modèle HA-Norm*

I	V1	V2	V3	R1			67	37,64
I	V1	V2	R1				43	61,80
I	V1	R1					13	69,10
I	R1						9	74,16
I	V1	V2	R1	V3			8	78,65

D *Modèle HA-Norm*

V1	V2	V3	I	R1			72	51,43
V1	V2	I	R1				30	72,86
V1	V2	I	V3	R1			19	86,43

Légende :
I : Infection — V1 : 1ère dose de vaccin — V2 : 2e dose de vaccin — V3 : 3e dose de vaccin
R1 : 1ère réinfection — R2 : 2e réinfection — R3 : 3e réinfection

Figure 5 - Séquences les plus fréquentes selon un seuil de 75% pour le modèle HD selon A) le groupe 1 et B) le groupe 2 et le modèle HA-Norm selon C) le groupe 1 et D) le groupe2.

Les deux séquences fréquentes suivantes, soit I-V1-V2-R1 et V1-V2-I-R1, concernaient 42 et 30 individus. La dernière séquence est celle des 20 patients dont l'immunité était naturelle plutôt qu'hybride, c'est-à-dire la séquence I-R1. Ces différentes séquences sont présentées à la Figure 6A.

Pour le second groupe du modèle HD-OM, mis à part la séquence partagée avec le groupe 1, les profils de réinfection supplémentaires étaient V1-V2-I-V3-R1 pour quatre individus, I-V1-V2-V3-R1-R2 pour trois personnes et enfin V1-V2-V3-R1-R2-R3 et V1-I-V2-V3-R1 pour chacun deux personnes. Dans ce groupe, les profils de réinfection les

plus fréquents comprennent tous les trois doses de vaccin et l'événement qui a précédé l'infection initiale est le 3e vaccin. Les combinaisons d'événements pour ce groupe sont présentées à la Figure 6B.

A

Événements							n	Fréquence cumulée (%)
1	2	3	4	5	6	7		
V1	V2	V3	I	R1			71	24,40
I	V1	V2	V3	R1			57	43,99
I	V1	V2	R1				43	58,76
V1	V2	I	R1				30	69,07
I	R1						20	75,94

B

I	V1	V2	V3	R1			10	37,04
V1	V2	I	V3	R1			4	51,85
I	V1	V2	V3	R1	R2		3	62,96
I	V1	V2	V3	R1	R2	R3	2	70,37
V1	I	V2	V3	R1			2	77,78

Légende :
I : Infection
V1 : 1ère dose de vaccin
V2 : 2e dose de vaccin
V3 : 3e dose de vaccin
R1 : 1ère réinfection
R2 : 2e réinfection
R3 : 3e réinfection

Figure 6 - Séquences les plus fréquentes selon un seuil de 75% pour le modèle HD-OM selon A) le groupe 1 et B) le groupe 2.

5.2 Caractéristiques sociodémographiques, état et habitudes du participant

Le Tableau 5 présente, pour les trois modèles, les caractéristiques des patients pour chacun des groupes obtenus. À titre de rappel, les regroupements ont été effectués en exploitant uniquement les séquences temporelles. Les variables de la présente section n'ont donc pas influencé la formation des groupes. Concernant l'âge moyen et l'indice de masse corporelle, aucune différence significative n'a été identifiée que ce soit pour les modèles HD, HA-Norm ou HD-OM. Cependant, il semblait y avoir plus d'hommes dans un groupe que dans l'autre, (allant d'environ 63% à 92% des hommes de la cohorte) avec une proportion plus importante de fumeurs, soit plus de 50%. Cette différence a été notée pour

les trois algorithmes bien que le modèle HD-OM avait toutefois la particularité de classer la totalité des fumeurs, ou anciens fumeurs, dans le même groupe.

Tableau 5 - Caractéristiques des patients de la cohorte selon les groupes des modèles HD, HA-Norm et HD-OM

Caractéristiques	Cohorte (n = 318)	HD		HA-Norm		HD-OM	
		Groupe 1 (n = 176)	Groupe 2 (n = 142)	Groupe 1 (n = 178)	Groupe 2 (n = 140)	Groupe 1 (n = 291)	Groupe 2 (n = 27)
Âge (année) : moyenne (écart-type)	43,0 (13,8)	42,4 (13,8)	43,8 (13,8)	42,5 (14,0)	43,7 (13,5)	43,0 (13,8)	43,0 (13,0)
18 à 34 *n*(%)	94 (29,6%)	58 (33,0%)	36 (25,4%)	59 (33,1%)	35 (25,0%)	88 (30,2%)	6 (22,2%)
35 à 44 *n*(%)	84 (26,4%)	43 (24,4%)	41 (28,9%)	43 (24,2%)	41 (29,3%)	75 (25,8%)	9 (33,3%)
45 à 64 *n*(%)	125 (39,3%)	68 (38,6%)	57 (40,1%)	68 (38,2%)	57 (40,7%)	114 (39,2%)	11 (40,7%)
65 ans et plus *n*(%)	15 (4,7%)	7 (4,0%)	8 (5,6%)	8 (4,5%)	7 (5,0%)	14 (4,8%)	1 (3,7%)
IMC : moyenne (écart-type)	26,9 (5,3)	26,9 (5,6)	28,0 (6,1)	26,8 (5,6)	28,1 (6,2)	27,4 (5,8)	27,5 (6,5)
Sexe à la naissance							
Femme *n*(%)	230 (72,3%)	121 (68,8%)	109 (76,8%)	122 (68,5%)	108 (77,1%)	210 (72,2%)	20 (74,1%)
Homme *n*(%)	88 (27,7%)	55 (31,3%)	33 (23,2%)	56 (31,5%)	32 (22,9%)	81 (27,8%)	7 (25,9%)
Statut tabagique							
Non-fumeur *n*(%)	230 (72,3%)	123 (69,9%)	107 (75,4%)	123 (69,1%)	107 (76,4%)	213 (73,2%)	17 (63,0%)
Fumeur *n*(%)	22 (6,9%)	15 (8,5%)	7 (4,9%)	15 (8,4%)	7 (5,0%)	22 (7,6%)	0 (0%)
Ex-fumeur *n*(%)	55 (17,3%)	28 (15,9%)	27 (19,0%)	29 (16,3%)	26 (18,6%)	46 (15,8%)	9 (33,3%)
Tabagisme passif *n*(%)	1 (0,3%)	1 (0,6%)	0 (0%)	1 (0,6%)	0 (0%)	1 (0,3%)	0 (0%)
Inconnu *n*(%)	10 (3,1%)	9 (5,1%)	1 (0,7%)	10 (5,6%)	0 (0%)	9 (3,1%)	1 (3,7%)
Cigarette électronique							
Oui *n*(%)	13 (4,1%)	8 (4,5%)	5 (3,5%)	8 (4,5%)	5 (3,6%)	12 (4,1%)	1 (3,7%)
Non *n*(%)	299 (94,0%)	163 (92,6%)	136 (95,8%)	164 (92,1%)	135 (96,4%)	274 (94,2%)	25 (92,6%)
Inconnu *n*(%)	6 (1,9%)	5 (2,8%)	1 (0,7%)	6 (3,4%)	0 (0%)	5 (1,7%)	1 (3,7%)
Usage de drogues							
Oui *n*(%)	30 (9,4%)	18 (10,2%)	12 (8,5%)	17 (9,6%)	13 (9,3%)	30 (10,3%)	0 (0%)
Non *n*(%)	282 (88,7%)	153 (86,9%)	129 (90,8%)	155 (87,1%)	127 (90,7%)	256 (88,0%)	26 (96,3%)
Inconnu *n*(%)	6 (1,9%)	5 (2,8%)	1 (0,7%)	6 (3,4%)	0 (0%)	5 (1,7%)	1 (3,7%)
Travailleur de la santé							
Oui *n*(%)	141 (44,3%)	55 (31,3%)	86 (60,6%)	55 (30,9%)	86 (61,4%)	136 (46,7%)	5 (18,5%)
Non *n*(%)	171 (53,8%)	116 (65,9%)	55 (38,7%)	118 (66,3%)	53 (37,9%)	149 (51,2%)	22 (81,5%)
Inconnu *n*(%)	6 (1,9%)	5 (2,8%)	1 (0,7%)	5 (2,8%)	1 (0,7%)	6 (2,1%)	0 (0%)
Lieu de résidence							
Domicile	315 (99,1%)	173 (98,3%)	142 (100%)	175 (98,3%)	140 (100%)	288 (99,0%)	27 (100%)
RPA	2 (0,6%)	2 (1,1%)	0 (0%)	2 (1,1%)	0 (0%)	2 (0,7%)	0 (0%)
CHSLD	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
RIRTF	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
En maison de chambres	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Inconnu n(%)	1 (0,3%)	1 (0,6%)	0 (0%)	1 (0,6%)	0 (0%)	1 (0,3%)	0 (0%)

Caractéristiques	Cohorte (n = 318)	HD		HA-Norm		HD-OM	
		Groupe 1 (n = 176)	Groupe 2 (n = 142)	Groupe 1 (n = 178)	Groupe 2 (n = 140)	Groupe 1 (n = 291)	Groupe 2 (n = 27)
Occupant du ménage							
Avec des membres de la famille	279 (87,7%)	160 (90,9%)	119 (83,8%)	161 (90,4%)	118 (84,3%)	255 (87,6%)	24 (88,9%)
Avec un aide-soignant.e	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Seul.e	31 (9,7%)	12 (6,8%)	19 (13,4%)	12 (6,7%)	19 (13,6%)	28 (9,6%)	3 (11,1%)
Colocataire	4 (1,3%)	1 (0,6%)	3 (2,1%)	2 (1,1%)	2 (1,4%)	4 (1,4%)	0 (0%)
Inconnu n(%)	4 (1,3%)	3 (1,7%)	1 (0,7%)	3 (1,7%)	1 (0,7%)	4 (1,45)	0 (0%)
Sévérité de la maladie							
Léger	299 (94,0%)	159 (90,3%)	140 (98,6%)	161 (90,4%)	138 (98,6%)	272 (99.0%)	27 (100%)
Modéré	11 (3,5%)	11 (6,3%)	0 (0%)	11 (6,2%)	0 (0%)	11 (3,8%)	0 (0%)
Sévère	8 (2,5%)	6 (6,3%)	2 (1,4%)	6 (3,4%)	2 (1,4%)	8 (2,7%)	0 (0%)
Décès	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Vague de l'infection							
1	22 (6,9%)	22 (12,5%)	0 (0%)	22 (12,4%)	0 (0%)	22 (7,6%)	0 (0%)
2	130 (40,9%)	130 (73,9%)	0 (0%)	130 (73,0%)	0 (0%)	113 (38,8%)	17 (63,0%)
3	22 (6,9%)	19 (10,8%)	3 (2,1%)	20 (11,2%)	2 (1,4%)	20 (6,9%)	2 (7,4%)
4	26 (8,2%)	5 (2,8%)	21 (14,8%)	6 (3,4%)	20 (14,3%)	24 (8,2%)	2 (7,4%)
5	56 (17,6%)	0 (0%)	56 (39,4%)	0 (0%)	56 (40,0%)	52 (17,9%)	4 (14,8%)
6	16 (5,0%)	0 (0%)	16 (11,3%)	0 (0%)	16 (11,4%)	15 (5,2%)	1 (3,7%)
7	46 (14,5%)	0 (0%)	46 (32,4%)	0 (0%)	46 (32,9%)	45 (15,5%)	1 (3,7%)
Vague de la réinfection							
1	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
2	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
3	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
4	1 (0,3%)	1 (0,6%)	0 (0%)	1 (0,6%)	0 (0%)	1 (0,3%)	0 (0%)
5	72 (22,6%)	65 (36,9%)	7 (4,9%)	65 (36,5%)	7 (5,0%)	64 (22,0%)	8 (29,6%)
6	43 (13,5%)	37 (21,0%)	6 (4,2%)	37 (20,8%)	6 (4,3%)	37 (12,7%)	6 (22,2%)
7	202 (63,5%)	73 (41,5%)	129 (90,8%)	75 (42,1%)	127 (90,7%)	189 (64,9%)	13 (48,1%)
Nombre de doses de vaccin lors de l'infection primaire : moyenne (écart-type)	1,08 (1,3)	0,0511 (0,22)	2,35 (0,88)	0,0618 (0,24)	2,36 (0,88)	1,11 (1,31)	0,704 (1,03)
Nombre de doses de vaccin lors de la première réinfection : moyenne (écart-type)	2,36 (0,88)	2,22 (0,87)	2,54 (0,85)	2,24 (0,86)	2,51 (0,88)	2.32 (0,90)	2.85 (0,36)
Nombre de doses de vaccin lors de la deuxième réinfection : moyenne (écart-type)	2,29 (0,86)	2,28 (0,79)	2,33 (1,21)	2,28 (0,79)	2,33 (1,21)	2.09 (0,92)	2.78 (0,44)

Caractéristiques	Cohorte (n = 318)	HD		HA-Norm		HD-OM	
		Groupe 1 (n = 176)	Groupe 2 (n = 142)	Groupe 1 (n = 178)	Groupe 2 (n = 140)	Groupe 1 (n = 291)	Groupe 2 (n = 27)
Nombre de doses de vaccin lors de la troisième réinfection : moyenne (écart-type)	2,50 (0,548)	2,50 (0,577)	2,50 (0,707)	2,50 (0,577)	2,50 (0,707)	2.00 (0)	2.75 (0.500)
Nombre de réinfections : moyenne (écart-type)	1,12 (0,375)	1,16 (0,429)	1,06 (0,286)	1,16 (0,427)	1,06 (0,288)	1.08 (0.300)	1.48 (0.753)

Au niveau de la consommation de cigarette électronique ainsi que de la consommation de drogues, les deux algorithmes basés sur les données interévénements n'ont pas révélé de différence entre les deux groupes alors que le modèle HD-OM, similairement au statut tabagique, a regroupé la majorité de ces consommateurs, voire leur totalité, dans le même groupe que les fumeurs. Au niveau de la profession, une proportion de travailleurs de la santé plus importante que celle de la cohorte a été associée dans le groupe 2 pour les modèles HD et HA-Norm (respectivement 60,6 % et 61,4% versus 44,3% pour la cohorte) indiquant ainsi qu'ils semblent partager des délais interévénements similaires. Le modèle HD-OM ne présente pas cette spécificité sachant que le groupe 2 contient une proportion de travailleurs de la santé moins importante que la proportion de la cohorte (18,5% contre 44,3%). En somme, les groupes, qui ont été formés en fonction d'événements temporels sans considérer les caractéristiques sociodémographiques, l'état et les habitudes des participants dans leur formation, mettent en évidence de façon assez marquée des regroupements de travailleurs de la santé et de fumeurs au sein des groupes.

5.3 Contexte de l'infection et de la réinfection

En ce qui a trait au contexte de l'infection, pour les modèles HD et HA-Norm, la répartition entre les vagues d'infection montre que la majorité des participants du groupe 1 l'ont été dans les vagues 1, 2 et 3 alors que les participants classés dans le groupe 2 ont plutôt été infectés lors des vagues 4 à 7. Pour ces mêmes modèles, la majorité des patients ayant été réinfectés dans les vagues 5 et 6 font partie du groupe 1 alors que la majorité des réinfections de la vague 7 sont dans le groupe 2. Le modèle HD-OM, quant à lui, ne discrimine pas les vagues d'infection et de réinfection aussi clairement. Pour l'infection, un fait saillant est que le second groupe contient une majorité de patients qui ont été infectés dans la 2^{e} vague de la pandémie (63%). Pour la réinfection, la proportion dans les groupes est similaire à celle de la cohorte.

Concernant le statut vaccinal lors de l'infection primaire, le groupe 1 du modèle HD compte une moyenne de 0,051 dose (médiane 0; min 0; max 1) pour un intervalle de confiance de 95% (IC 95%) de 0,018-0,084 comparativement à 2,35 doses (médiane : 3, min 0, max 3, IC 95% : 2,203-2,497) pour le groupe 2. Le modèle HA-Norm présente des résultats similaires avec une moyenne de 0,062 dose (médiane 0; min 0; max 1, IC0,026-

0,098) pour le groupe 1 comparativement à 2,36 doses (médiane : 3, min 0, max 3, IC 95% : 2,214-2,506) pour le groupe 2. Cette différence est statistiquement significative (p<0,001) pour les deux modèles. Pour le modèle HD-OM, la différence de moyenne pour cette même variable n'est pas concluante (p=0,161) avec le groupe 1 qui comptait une moyenne de 1,11 dose (IC 95% : 0,958-1,262) comparativement à 0,704 dose (IC 95% : 0,296-1,112) pour le groupe 2.

Pour le statut vaccinal lors de la réinfection, les deux modèles basés sur les délais interévénements présentent des valeurs p inférieures à 0,001. Pour le HD, le nombre de doses du groupe 1 compte une moyenne de 2,22 doses (IC 95% : 2,09-2,35) contre une moyenne de 2,54 (IC 95% : 2,399-2,681) alors que pour le HA-Norm, le nombre de doses du groupe 1 compte une moyenne de 2,24 doses (IC 95% : 2,113-2,367) contre une moyenne de 2,54 (IC 95% : 2,364-2,657). Le modèle sur les séquences d'événements, quant à lui, présente pour le groupe 1 une moyenne de 2,32 doses (IC 95% : 3,217-3,423) contre une moyenne de 2,85 (IC 95% : 2,707-2,993) avec une valeur p de 0,001. Pour les réinfections subséquentes, il n'y a pas d'évidence statistique que les moyennes des groupes sont différentes, et ce, pour les trois modèles.

Finalement, pour le nombre moyen de réinfections survenues, il existe une différence du point de vue statistique pour les trois modèles, bien que cette dernière n'apporte pas d'élément significatif du point de vue scientifique. Avec une valeur p de 0,004, le modèle HD avait, pour son premier groupe, une moyenne de 1,16 (IC 95% : 1,096-1,224) contre 1,06 (IC 95% : 1,013-1,107) pour le deuxième groupe alors que ceux du modèle HA-Norm (p<0,003) avait une moyenne identique, pour un intervalle de confiance respectif de 1,097-1,223 et 1,012-1,108. Le modèle HD-OM avait, pour sa part, une valeur p inférieur à 0,001 et une moyenne de 1,08 (IC 95% : 1,045-1,115) pour le groupe 1 contre 1,48 (IC 95% : 1,182-1,778) pour le groupe 2.

En somme, pour les modèles basés sur les délais interévénements, la vague de l'infection et de la réinfection, bien que non utilisée comme intrant dans l'algorithme pour la formation des groupes, distinguait les deux groupes en séparant les individus selon la temporalité de leur épisode de la maladie. De même, le statut vaccinal lors de l'infection

apparaissait également comme distinctif classant les individus en moyenne moins vaccinés ensemble. Ces individus s'avéraient d'ailleurs être dans le groupe ayant été infectés dans les premières vagues de la pandémie. Pour le modèle basé sur la séquence, c'est plutôt le nombre de réinfections qui apparaissait distinguer les deux groupes.

5.4 Délais interévénements

Relativement aux délais entre les événements d'intérêt, le Tableau 6 présente leur moyenne alors que le Tableau 7 en présente les intervalles de confiance à 95% ainsi que la valeur p de chacun des modèles. Pour rappel, les variables des délais interévénements ont, en plus de servir à la description des groupes, étaient utilisées dans la construction des modèles HD et HA-Norm.

Pour ces derniers, il n'y a pas de différence statistiquement significative entre la première et deuxième dose de vaccin. Il en est de même pour la première dose et la troisième dose. Pour le modèle HD-OM, les deux groupes présentent une différence de moyenne statistiquement significative entre la première et la deuxième dose. Cependant, il s'agit, pour ce modèle, de la seule variable de délais impliquant la première dose de vaccin qui s'avérait significative en termes de différence de moyenne entre les groupes. Pour les autres modèles, la moyenne des deux groupes concernant le délai entre la première dose de vaccin et les événements d'infection, de première réinfection et de deuxième réinfection était différente, mais ce n'était pas le cas avec la troisième réinfection.

Pour les combinaisons de délais impliquant la deuxième dose de vaccin non décrites précédemment, la totalité des variables ressortait statistiquement significative en termes de différence de moyenne entre les deux groupes des modèles HD et HA-Norm. À l'instar des variables de délais incluant la première dose de vaccin, aucun délai entre la deuxième dose et les autres événements d'intérêt n'était différent d'un point de vue statistique pour le modèle HD-OM.

Pour ce qui est des délais impliquant la troisième dose de vaccin qui n'ont pas été présentés ci-dessus, la moyenne des deux groupes quant aux délais avec l'infection primaire ainsi qu'avec la première réinfection se révélaient différents, et ce, pour les trois

modèles. Pour les modèles HD et HA-Norm, il en est de même pour le délai entre la troisième dose et la deuxième réinfection. Ce n'est toutefois pas le cas pour le modèle HD-OM pour lequel il n'y a pas suffisamment d'évidences statistiques pour considérer ce délai comme différent.

En ce qui concerne les délais combinant l'infection primaire et les événements d'intérêt qui n'ont pas été énoncés précédemment, à savoir avec les trois réinfections, les deux groupes des modèles HD et HA-Norm présentent une différence statistique entre la moyenne des délais entre l'infection primaire et les trois réinfections. À contrario, aucun de ces trois délais n'est différent d'un point de vue statistique pour les deux groupes du modèle HD-OM.

Finalement, pour les événements restants, à savoir les délais entre les réinfections entre elles, il existait suffisant d'évidences statistiques pour considérer la moyenne du délai entre la première réinfection et la seconde réinfection comme différente pour les deux groupes des modèles HD et HA-Norm. Pour le délai entre les autres associations d'événements, les délais n'étaient pas différents d'un point de vue statistique. Quant au modèle HD-OM, la totalité des délais moyens pour les réinfections entre elles n'était pas statistiquement différents entre les deux groupes.

Pour résumer, la majorité des délais, à l'exception de ceux entre le premier vaccin et les doses suivantes ainsi que les délais impliquant les réinfections subséquentes à la première, étaient distinctifs pour les groupes des modèles construits sur ces mêmes délais. Pour les groupes issus des séquences d'événements, c'est plutôt la situation inverse à savoir que la très grande majorité des délais ne sont pas distinctifs entre les regroupements sauf pour le délai entre deux premières doses de vaccins ainsi que les délais entre la 3^{e} dose et l'infection primaire et la première réinfection.

Tableau 6 - Délais interévénements des patients de la cohorte selon les groupes des modèles HD, HA-Norm et HD-OM

Délai interévénements (jours) : moyenne (écart-type)	**Cohorte (n = 318)**	**HD**		**HA-Norm**		**HD-OM**	
		Groupe 1 (n = 176)	**Groupe 2 (n = 142)**	**Groupe 1 (n = 178)**	**Groupe 2 (n = 140)**	**Groupe 1 (n = 291)**	**Groupe 2 (n = 27)**
Vaccin1-Vaccin2	103 (63,6)	116 (78,8)	87.6 (33,6)	116 (79,1)	87,3 (33,9)	104 (63,8)	88.5 (61,70)
Vaccin1-Vaccin3	298 (67,8)	293 (75,3)	303 (59,3)	295 (74,6)	300 (61,0)	298 (67,3)	297 (71,70)
Vaccin1-Infection	707 (267)	-143 (103)	351 (114)	-146 (101)	343 (123)	78,0 (274)	-2.33 (178)
Vaccin1-Réinfection	463 (172)	364 (125)	594 (132)	362 (124)	592 (133)	469 (175)	412 (134)
Vaccin1-Réinfection2	523 (151)	478 (99,0)	747 (177)	478 (99,0)	747 (177)	537 (171)	489 (91,10)
Vaccin1-Réinfection3	583 (113)	553 (133)	643 (21,9)	553 (133)	643 (21,9)	579 (206)	585 (85,20)
Vaccin2-Vaccin3	202 (60,1)	191 (66,5)	213 (50,9)	192 (66,7)	211 (52,0)	201 (59,5)	207 (65,20)
Vaccin2-Infection	-18,6 (285)	-257 (125)	263 (108)	-261 (123)	256 (117)	-10,8 (293)	-90.9 (183)
Vaccin2-Réinfection	373 (182)	260 (138)	507 (129)	257 (136)	504 (131)	379 (184)	323 (148)
Vaccin2-Réinfection2	407 (190)	348 (147)	656 (150)	348 (147)	656 (150)	413 (226)	397 (102)
Vaccin2-Réinfection3	467 (101)	410 (62,7)	581 (18,4)	410 (62,7)	581 (18,4)	374 (4,95)	514 (91,40)
Vaccin3-Infection	-169 (269)	-416 (108)	74.3 (112)	-423 (101)	66,2 (120)	-148 (275)	-309 (173)
Vaccin3-Réinfection	213 (166)	110 (133)	315 (129)	105 (129)	313 (131)	229 (162)	112 (157)
Vaccin3-Réinfection2	245 (169)	171 (98,0)	486 (111)	171 (98,0)	486 (111)	316 (187)	164 (107)
Vaccin3-Réinfection3	315 (92,1)	275 (84,9)	396 (--)	275 (84,9)	396 (--)	--	315 (92,10)
Infection-Réinfection	391 (177)	510 (130)	240 (92,4)	510 (130)	244 (100)	389 (178)	414 (169)
Infection-Réinfection2	564 (154)	622 (108)	326 (44,5)	622 (108)	326 (44,5)	580 (155)	526 (154)
Infection-Réinfection3	583 (170)	686 (75,2)	376 (7,07)	686 (75,2)	376 (7,07)	706 (80,6)	521 (176)
Réinfection-Réinfection2	175 (86,7)	182 (91,9)	145 (56,5)	182 (91,9)	145 (56,5)	185 (83,4)	152 (95,40)
Réinfection-Réinfection3	210 (69,6)	222 (78,4)	186 (62,9)	222 (78,4)	186 (62,9)	253 (93,3)	189 (57,60)
Réinfection2-Réinfection3	96,5 (69,6)	102 (88,4)	85.5 (19,1)	102 (88,4)	85,5 (19,1)	92,5 (108)	98.5 (64,42)

Tableau 7 - Intervalles de confiance et valeur p des délais interévénements pour les modèles HD, HA-Norm et HD-OM

	HD			HA-Norm			HD-OM		
	Intervalle de confiance 95%		Valeur p	Intervalle de confiance 95%		Valeur p	Intervalle de confiance 95%		Valeur p
Délais (en jours)	**Groupe 1**	**Groupe 2**		**Groupe 1**	**Groupe 2**		**Groupe 1**	**Groupe 2**	
Vaccin1-Vaccin2	103,52, 129,32	81,48, 93,2	0,145 NS	102,93, 128,34	81,7, 93,46	0,180 NS	96,31, 112,17	64,09, 112,94	0,028 *
Vaccin1-Vaccin3	279,97, 310,68	287,97, 312,19	0,146 NS	277,6, 308,13	290,71, 314,61	0,559 NS	287,66, 308,24	267,8, 325,74	0,370 NS
Vaccin1-Infection	-161,08, -130,06	321,73, 364,36	< 0,001 ***	-158,21, -126,91	330,59, 370,51	< 0,001 ***	45,13, 110,88	-72,72, 68,05	0,240 NS
Vaccin1-Réinfection	342,49, 380,73	568,68, 614,79	< 0,001 ***	344,53, 382,72	571,35, 617,61	< 0,001 ***	447,69, 489,62	358,78, 464,48	0,100 NS
Vaccin1-Réinfection2	436,72, 518,48	527,82, 966,58	< 0,001 ***	436,72, 518,48	527,82, 966,58	< 0,001 ***	459,19, 614,53	419,09, 559,13	0,438 NS
Vaccin1-Réinfection3	341,49, 764,51	445,55, 839,45	0,422 NS	341,49, 764,51	445,55, 839,45	0,422 NS	-1 270,25, 2 427,25	449,45, 720,55	0,956 NS
Vaccin2-Vaccin3	178,57, 206,04	200,87, 221,49	< 0,001 ***	177,19, 204,13	203,13, 223,66	< 0,001 ***	192,22, 210,38	180,78, 233,45	0,741 NS
Vaccin2-Infection	-281,27, -241,46	235,57, 275,85	< 0,001 ***	-277,3, -237,19	244,15, 281,79	< 0,001 ***	-47,17, 25,49	-163,24, -18,46	0,315 NS
Vaccin2-Réinfection	235,17, 279,21	481,76, 527,03	< 0,001 ***	237,83, 282,14	484,37, 529,42	< 0,001 ***	355,76, 401,51	264,59, 381,64	0,981 NS
Vaccin2-Réinfection2	281,5, 414,89	470,04, 841,56	< 0,001 ***	281,5, 414,89	470,04, 841,56	< 0,001 ***	296,71, 528,94	318,49, 475,51	0,845 NS
Vaccin2-Réinfection3	310,04, 509,46	415,82, 746,18	0,023 *	310,04, 509,46	415,82, 746,18	0,023 *	329,03, 417,97	368,07, 658,94	0,111 NS
Vaccin3-Infection	-443,83, -402,04	42,48, 89,98	< 0,001 ***	-437,6, -393,94	51,68, 96,86	< 0,001 ***	-189,85, -105,78	-378,57, -238,74	0,008 **
Vaccin3-Réinfection	78,75, 131,86	287,3, 339,16	< 0,001 ***	83,04, 137,1	289,02, 340,86	< 0,001 ***	203,95, 253,58	48,5, 175,5	0,001 **
Vaccin3-Réinfection2	111,35, 229,73	309,55, 662,45	< 0,001 ***	111,35, 229,73	309,55, 662,45	< 0,001 ***	172,79, 459,66	74,88, 253,87	0,062 NS
Vaccin3-Réinfection3	-487,37, 1 037,37	--	--	-487,37, 1 037,37	--	--	368,69, 409,72	86,57, 544,09	--

	HD			HA-Norm			HD-OM		
	Intervalle de confiance 95%		Valeur p	Intervalle de confiance 95%		Valeur p	Intervalle de confiance 95%		Valeur p
Délais (en jours)	**Groupe 1**	**Groupe 2**		**Groupe 1**	**Groupe 2**		**Groupe 1**	**Groupe 2**	
Infection-Réinfection	490,76, 529,39	227,48, 260,73	< 0,001 ***	491,03, 529,61	224,56, 255,43	< 0,001 ***	511,37, 648,72	347,26, 480,66	0,493 NS
Infection-Réinfection2	577,26, 666,1	278,82, 372,18	< 0,001 ***	577,26, 666,1	278,82, 372,18	< 0,001 ***	-18,25, 1430,25	407,57, 644,43	0,384 NS
Infection-Réinfection3	566,03, 805,47	312,47, 439,53	0,003 **	566,03, 805,47	312,47, 439,53	0,004 **	147,52, 221,48	241,03, 800,47	0,246 NS
Réinfection-Réinfection2	144,47, 220,34	85,26, 203,74	< 0,001 ***	144,47, 220,34	85,26, 203,74	< 0,001 ***	-585,61, 1091,61	78,68, 225,32	0,384 NS
Réinfection-Réinfection3	97,52, 346,99	-379,93, 750,93	0,601 NS	97,52, 346,99	-379,93, 750,93	0,601 NS	-879,53, 1064,53	96,92, 280,08	0,498 NS
Réinfection2-Réinfection3	-38,73, 242,73	-86,03, 257,03	0,742 NS	-38,73, 242,73	-86,03, 257,03	0,742 NS	96,31, 112,17	-3,97, 200,97	0,933 NS

5.5 Analyses de sensibilité

Constance des attributs distinctifs. Les caractéristiques des groupes obtenus par les différentes méthodes de classification ont permis de mettre en évidence des attributs spécifiques. En les comparant pour chaque méthode, les résultats révèlent des attributs similaires pour une grande proportion des variables d'intérêts. Cependant, dans certains cas, une méthode met en lumière des attributs différents des autres méthodes par le biais de constats supplémentaires ou, au contraire, par l'absence de certains attributs. Le Tableau 8 dresse un sommaire de ces différences pour les variables d'intérêt visées.

Tableau 8 - Sommaire des différences dans la caractérisation pour certaines variables d'intérêts pour chacun des modèles

Variables d'intérêt	Modèles de regroupement			
	HD	HA-Norm	HD-OM	HA-DTW
Statut tabagique	•	•	+	•
Cigarette électronique	•	•	+	•
Usage de drogues	•	•	+	•
Travailleur de la santé	•	•	+	•
Occupant du ménage	•	•	+	•
Vague de l'infection	•	•	-	•
Vague de la réinfection	•	•	•	+
Nombre de doses de vaccin lors de l'infection primaire	•	•	-	+
Nombre de doses de vaccin lors de la réinfection	•	•	•	+
Nombre de doses de vaccin lors de la 2e réinfection	•	•	•	+
Nombre de réinfections	•	•	•	+

Légende
- \- Des attributs distinctifs sont manquants par rapport aux autres modèles de regroupement
- • Les attributs distinctifs sont identiques ou similaires parmi les modèles de regroupement
- \+ Des attributs distinctifs supplémentaires sont présents par rapport aux autres modèles de regroupement

Les modèles HD et HA-Norm ne présente, entre eux, que des différences mineures dans les effectifs des groupes. Ces différences, majoritairement de +/- 1 individus entre les groupes, n'entraîne ainsi pas de conclusions divergentes dans la caractérisation des groupes. Le modèle HD-OM, quant à lui, présente des différences sur certaines variables dans le sens où il regroupe, en totalité ou en majorité, les patients qui consomment tabac, cigarette électronique ou drogues, au sein d'un même groupe. De même, les travailleurs

de la santé sont, pour la plupart, dans ce même groupe. Bien que cet aspect se distingue des autres modèles, le HD-OM ne présente pas, contrairement aux autres modèles, de distinction particulière dans vagues et dans le nombre de doses de vaccins de l'infection primaire. Le modèle HA-DTW, pour sa part, apporte des nuances par rapport aux autres modèles au niveau de la vaccination. Effectivement, en ce qui concerne l'infection primaire, le modèle met l'emphase sur le nombre de doses dans l'assignation aux groupes sachant qu'un groupe contient que des individus sans doses, un groupe contient une majorité d'individus avec 1 dose, un autre 2 doses et un dernier avec 3 doses. Dans un ordre d'idée similaire, la distribution d'un des groupes suggère qu'une proportion importante d'individus avec un nombre de 0 ou 1 dose lors de leur réinfection sont regroupés ensemble. Il regroupe également les individus avec 0 ou 1 doses lors de leur 2e réinfection ensemble. De ce fait, deux groupes ne contiennent que des individus ayant été réinfectés une fois alors qu'un groupe n'a aucun individu ayant été réinfecté 3 fois.

Stabilité des groupes. L'analyse de la stabilité des groupes entre les modèles a révélé qu'en considérant les modèles sélectionnées, 22,6 % des paires d'individus se retrouvaient groupés ensemble par quatre des modèles, 24,1 % par trois des modèles, 5 % par deux des modèles uniquement et 41,1 % par un seul des modèles. A contrario, 7,2% des paires d'individus ne se retrouvent jamais dans le même groupe, peu importe le modèle. Ainsi, pour 29,8% des paires d'individus, les quatre modèles assignent les individus de la même façon à savoir toujours ensemble ou jamais ensemble. La Figure 7 présente une charte de chaleur de la fréquence où chaque paire d'individus se retrouve dans le même groupe.

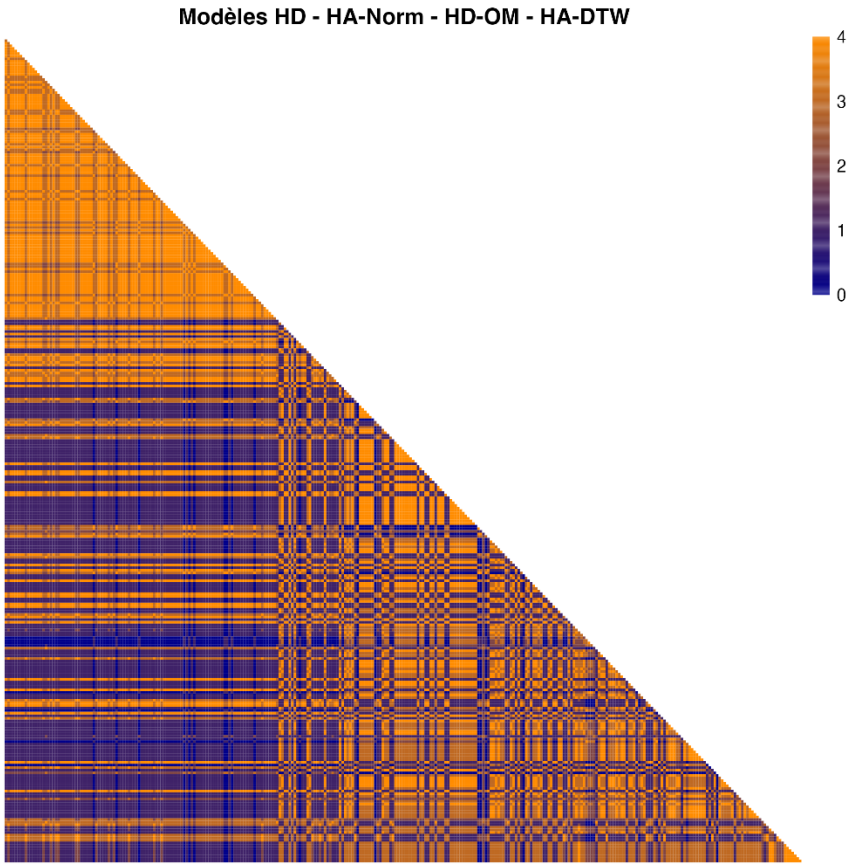


Figure 7 - Carte de chaleur de la stabilité des classements présentant le nombre d'occurrences où chaque paire d'individus est groupée ensemble par les modèles

En retirant à tour de rôle un modèle de l'analyse, le taux d'assignation identique des paires d'individus demeure relativement stable lors du retrait des modèles HD et HA-Norm (respectivement 30,4% et 30,2%), mais varie de manière plus significative lors du retrait des modèles HD-OM et HA-DTW (respectivement 75% et 48,7%). La Figure 8 reprend la carte de chaleur de la Figure 7, mais pour chacune des itérations.

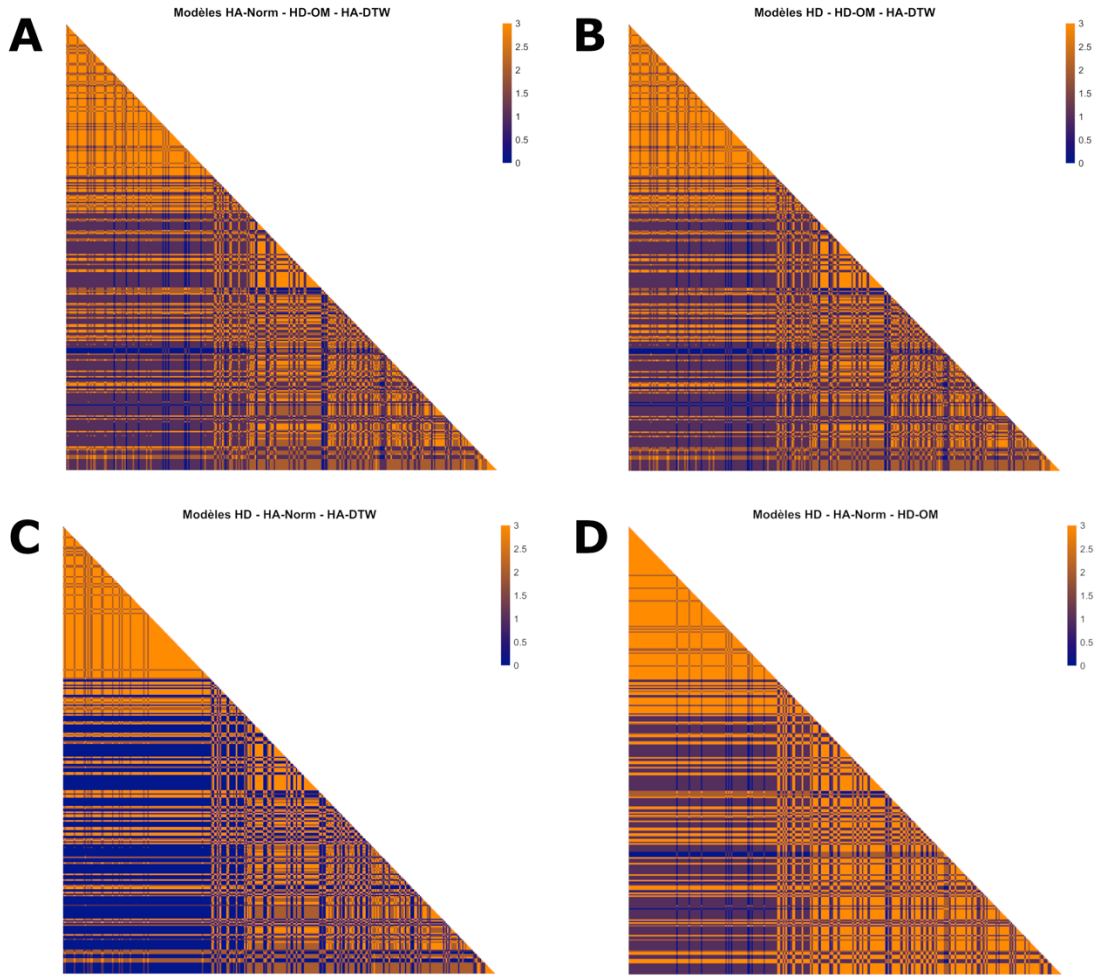


Figure 8 - Cartes de chaleur de la stabilité des classements présentant le nombre d'occurrences où chaque paire d'individus est groupée ensemble par les modèles, mais en retirant successivement un modèle

Enfin, en comparant les modèles à tour de rôle les uns avec les autres, les modèles HA-Norm et HD obtiennent un taux de classification identique de 97,5% des paires d'individus. Le taux de classification, qui correspond ici à la proportion de paires d'individus classée dans le même groupe par les deux modèles par rapport au nombre de paires d'individus totales, s'avère similaire pour ces modèles lorsqu'ils sont comparés aux autres : tous deux obtiennent 50% avec le modèle HD-OM et respectivement 76,1% et 76,5% avec le modèle HA-DTW. Les modèles HD-OM et HA-DTW, quant à eux, classent 34,3% des paires d'individus de la même façon. L'Annexe 6 présente des cartes de chaleur de chacune des comparaisons de modèles entre eux alors que le Tableau 9 présente les résultats de toutes les analyses réalisées.

Ainsi, les résultats des deux modèles basés sur les délais interévénements, bien que leurs algorithmes de classification étaient différents, sont très similaires et, conséquemment, très stables entre eux. En revanche, le modèle impliquant l'OM produit des résultats différents des trois autres modèles et pourrait ainsi être considéré moins stables que les autres. Finalement, le modèle basé sur une matrice de dissimilarité produite avec l'algorithme *Dynamic Time Warping* est plus similaire aux modèles découlant des délais interévénements que le modèle HD-OM, tout en présentant certaines distinctions dans l'assignation des individus aux groupes. Avec un taux de classification identique des paires d'individus d'environ 76% avec les deux modèles HD et HA-Norm, le modèle HA-DTW présentait une stabilité intéressante pour un nombre de groupes d'individus plus important (5 groupes versus 2 groupes), et par le fait même, un potentiel de caractérisation plus nuancée.

Tableau 9 - Paires de patients assignés au même groupe présentés selon le nombre de modèles ayant réalisé l'assignation au groupe

Scénario d'analyse	**Paires de patients ayant été classées dans le même groupe selon le nombre de modèles les ayant assignés au même groupe (n = 50 403)**				
	0	**1**	**2**	**3**	**4**
Tous les modèles n(%)	3 607 (7,2%)	20 729 (41,1%)	2 515 (5,0%)	12 159 (24,1%)	11 393 (22,6%)
Retrait d'un des modèles					
Excluant HD n(%)	3 771 (7,5%)	22 399 (44,4%)	12 701 (25,2%)	11 532 (22,9%)	--
Excluant HA-Norm n(%)	3 782 (7,5%)	22 341 (44,3%)	12 856 (25,5%)	11 424 (22,7%)	--
Excluant HD-OM n(%)	23 989 (47,6%)	1 398 (2,8%)	11 184 (22,2%)	13 832 (27,4%)	--
Excluant HA-DTW n(%)	3 615 (7,2%)	21 079 (41,8%)	4 766 (9,5%)	20 943 (41,5%)	--
Comparaison des modèles paire par paire					
HD – HA-Norm n(%)	24 328 (48,3%)	1 256 (2,5%)	24 819 (49,2%)	--	--
HD – HD-OM n(%)	3 816 (7,6%)	25 217 (50,0%)	21 370 (42,4%)	--	--
HD– HA-DTW n(%)	24 488 (48,6%)	12 051 (23,9%)	13 864 (27,5%)	--	--
HA-Norm – HD-OM n(%)	3 780 (7,5%)	25 217 (50,0%)	21 406 (42,5%)	--	--
HA-Norm – HA-DTW n(%)	24 549 (48,7%)	11 857 (23,5%)	13 997 (27,8%)	--	--
HD-OM – HA-DTW n(%)	5 383 (10,7%)	33 126 (65,7%)	11 894 (23,6%)	--	--

Chapitre 6 Discussion

Ce projet avait pour objectif principal d'exploiter les données de la Biobanque québécoise de la COVID-19 à l'aide de techniques d'apprentissage automatique afin d'identifier et de caractériser les différents profils de réinfection au SRAS-CoV-2 pour étudier l'immunité hybride. Spécifiquement, le projet visait à regrouper les individus selon leur similarité (« *pattern* ») de séquence temporelle de vaccination, d'infection et de réinfection pour caractériser les groupes ainsi obtenus en termes de facteurs sociodémographiques et cliniques. Un sous-objectif était d'appliquer et comparer différentes méthodologies d'apprentissage automatique afin d'identifier la ou les plus pertinentes pour étudier ce type de problématique en conservant un équilibre entre la performance de l'algorithme et l'interprétabilité.

Au préalable, les données de la BQC19 ont été nettoyées et préparées afin de les optimiser pour les analyses de données temporelles. Cette étape a d'ailleurs permis la création d'un script de nettoyage et de transformation autoportant pouvant être utilisé de manière systématique sur les jeux de données concernées permettant ainsi à la BQC19 d'en bénéficier et de le rendre disponible pour faciliter l'exploitation des données dans d'autres projets de recherche. En effet, ce dernier, étant optimisé pour traiter l'ensemble de la population de la BQC19 et non uniquement la cohorte de l'étude, constitue une ressource particulièrement intéressante à partager.

Concernant la cohorte spécifique à l'étude, les critères d'inclusion exigeaient que les participants soient majeurs, qu'ils disposent d'une infection primaire et d'une réinfection, toutes deux documentées avec une date. Des 6 271 individus de la BQC19, 318 ont finalement été inclus dans l'étude, ce qui représente une proportion relativement faible, le critère concernant la réinfection pleinement documentée étant celui ayant le plus réduit la taille de l'échantillon.

Au terme du projet, quatre modèles ont été retenus sur la base du coefficient de silhouette moyen. Pour les quatre modèles, les groupes ont été formés avec une approche « *data-driven* » où, à l'aide de techniques d'apprentissages non-supervisées, les groupes

ont été formés en exploitant exclusivement les séquences temporelles des épisodes de COVID-19 des patients. Ainsi, les groupes ont principalement été caractérisés avec des variables « neutre » quant à leur contribution dans la constitution des groupes.

6.1 Caractéristiques

En premier lieu, l'étude a démontré qu'en général, il n'y avait pas de distinctions significatives entre les groupes au niveau des variables sociodémographiques, mis à part quelques différences de proportions entre certains groupes par rapport aux proportions de la cohorte. En effet, certains groupes avaient la particularité de comprendre davantage de fumeurs. Un des modèles les regroupaient même tous au sein d'un seul groupe avec les anciens fumeurs, les utilisateurs de cigarettes électroniques ainsi que les consommateurs de drogues. Pour ce dernier, la prudence est toutefois de mise quant aux interprétations possibles étant donné l'important débalancement en termes de nombre d'individus entre les groupes (291 versus 27). Ensuite, la proportion plus élevée de travailleurs de la santé (TdeS) dans certains groupes, et ce, pour la majorité des algorithmes, suggère une similarité dans la séquence temporelle des travailleurs de la santé. Dans certains cas, ces derniers étaient regroupés avec les fumeurs, ce qui pourraient laisser penser que certains groupes pouvaient mettre en évidence des individus plus vulnérables.

Ensuite, le projet mettait en lumière le rôle majeur de la temporalité des événements. En effet, sans égard à l'algorithme, une tendance se dégage quant au regroupement des individus en fonction du moment où leur infection primaire est survenue, à savoir dans les trois premières vagues de la pandémie ou dans les vagues subséquentes. Les événements de vaccination suivaient une logique inverse dans le sens où les groupes infectés plus tardivement étaient en moyenne les plus hâtivement vaccinés. Ces derniers correspondaient également aux groupes où la proportion de travailleurs de la santé était plus importante. Ce constat, impliquant que les travailleurs de la santé de la cohorte ont été infectés tardivement (59,6% des TdeS), diffère des résultats de Carazo *et al.* (2023) où, pour environ la même période, approximativement 20,7% des travailleurs de la santé l'ont été. Également, les différents profils de réinfection démontraient que très peu de réinfection étaient survenues avant la fin de 2021 sachant qu'elles apparaissaient principalement à compter de la 5^{e} vague. Cette dernière correspondait à l'apparition du

variant Omicron dans la population, jugé plus transmissible (Groupe de travail sur l'immunité face à la COVID-19, 2024). Ces résultats sont notamment cohérents à ceux de Keeling (2023) et de Chen *et al.* (2024) qui soulevaient entre autres la prépondérance des réinfections avec l'apparition du variant Omicron. Globalement, il a été possible de constater, lors de l'analyse des séquences d'événements des groupes, que les séquences impliquant des vaccins **avant** l'événement d'infection primaire ou secondaire, même si des doses étaient manquantes pour considérer l'immunisation complète, semblaient retarder l'événement d'infection subséquent. Ce fait semblait cohérent avec les politiques de vaccination ayant eu cours pendant la pandémie pour cette population à risque et en contact accru avec le virus. Les résultats suggéraient donc que ces politiques ont eu un impact positif étant donné que l'infection primaire est survenue plus tardivement pendant la pandémie.

Dans un ordre d'idée similaire, l'analyse de l'un des groupes (le groupe 1), issu du modèle généré avec le DTW, présentait un élément intéressant quant à l'efficacité potentielle de la vaccination. En effet, à la suite du second vaccin, la trajectoire des individus se séparaient en deux, certains vers la réinfection alors que les autres s'orientaient vers le troisième vaccin. Cette séparation survenait dans un délai médian presqu'identique (185 et 181 jours), ce qui pourrait suggérer que la politique d'administration de la troisième dose était relativement synchronisée avec un affaiblissement de l'immunité. Ce délai s'inscrit en adéquation avec les résultats de Asamoah-Boaheng *et al.* (2023) à l'effet que le niveau d'anticorps décroit avec une demi-vie de 94 jours et un plateau à 294 jours. Bien que ces résultats portent sur les vaccins à ARN messager et que le type de vaccin n'était pas une variable du présent projet, les résultats demeurent cohérents. Aussi, le 3^{e} vaccin semblait avoir retardé la réinfection de 118 jours (médiane) par rapport aux patients n'ayant reçus que 2 doses. Cela laisse penser qu'une 3e dose administrée plus hâtivement aurait potentiellement pu éviter des réinfections. Le groupe 3 de ce même modèle avait une séparation similaire entre l'événement de réinfection et le troisième vaccin. Ce dernier, dont la séquence avant la séparation était V1-I-V2, comparativement à celle du premier groupe I-V1-V2, avait un temps médian avant la réinfection plus grand (215 vs. 185 jours). De même, en comparant le temps médian entre la 2^{e} dose de vaccin et la réinfection en passant par la 3^{e} dose de vaccin, le groupe 3 avait un temps plus long (348

jours versus 299 jours). Cela pourrait laisser penser qu'en termes d'immunité hybride, le fait d'être infecté entre deux doses de vaccins pourraient offrir une immunité légèrement plus durable, mais en raison de la taille du 3e groupe, il est difficile de tirer une telle conclusion.

L'étude a également mis en évidence un groupe de patients qui n'ont pas été vaccinés et, conséquemment, n'ont pas atteint l'immunité hybride. Certains individus n'avaient également reçu qu'un seul vaccin, impliquant alors qu'ils n'avaient pas pleinement atteint une immunité hybride, sachant qu'une l'immunité vaccinale complète requiert deux doses. Ainsi, les individus avec une immunité hybride partielle ou inexistante ont été regroupés ensemble par les algorithmes exploitant la séquence des événements et leur temporalité. Bien qu'intéressant, ce constat devait être nuancé en raison de la variance de la durée du suivi à l'intérieur des groupes. En effet, certains individus ne pourraient tout simplement pas avoir eu le temps de suivi requis pour être avoir été pleinement vacciné. Malgré cette réserve, il s'avère intéressant de mentionner que le regroupement de ces participants par l'algorithme dans une approche « *data-driven* » supporte le constat de Sanchez-de Prada *et al.* (2024) à savoir qu'il n'y a pas de différence significative entre les individus vaccinés une seule fois dans les cinq mois **après** l'infection et ceux qui n'ont pas été vaccinés du tout. Cela mentionné, des analyses complémentaires seraient pertinentes afin d'évaluer l'impact de la différence dans le temps de suivi des participants.

En somme, en termes de caractérisation, les résultats suggéraient que la vaccination a un effet positif en retardant l'infection ou la réinfection. Ils montraient également que la temporalité des événements a grandement influencé la formation des groupes par l'algorithme dans le sens où les infections primaires et les réinfections sont répartis selon une progression temporelle, allant des groupes avec les infections les plus précoces aux groupes avec les plus tardives.

6.2 Méthodes

Outre les objectifs de regrouper et caractériser les individus en fonction de leur profil de réinfection, le projet avait également pour but secondaire l'exploration des algorithmes d'apprentissage automatique et des mesures de dissimilarité les plus adaptés, c'est-à-dire

qui respectaient un équilibre entre la performance et l'interprétabilité, pour réaliser des regroupements en présence de « *patterns* » temporels complexes issus des données de la BQC19.

Différentes combinaisons de mesures de dissimilarité, d'algorithmes et des sources de données ont été testés et, parmi ces 368 possibilités, 4 modèles ont été retenus pour les analyses, et ce, sur la base du coefficient de silhouette moyen. Considérant le nombre de combinaisons et des modèles différents impliqués, la stabilité des modèles a dû être évaluée.

D'abord, cette étape du projet a révélé que les modèles sélectionnés étaient relativement stables, à l'exception de celui exploitant l'*Optimal Matching* (OM). En effet, ce dernier s'est révélé différents des autres modèles, ce qui est cohérent avec son coefficient de silhouette moyen qui était le plus faible des quatre modèles. Ce constat s'avérait particulièrement surprenant sachant que l'une des utilités principales de cet algorithme est justement l'analyse de séquences, dont celles temporelles. Dans une perspective du projet, il pourrait être intéressant d'évaluer si la modification des paramètres de génération de la matrice de coûts pourrait apporter des résultats différents. Effectivement, dans le cadre du projet, les différentes opérations d'alignement des événements dans les séquences (ajout / retrait / substitution) avaient un poids identique, mais une pondération différente serait intéressante à évaluer. De même, parmi les scénarios d'analyse du projet, il avait été envisagé d'utiliser l'algorithme OM sur les données en fixant l'infection de tous les patients comme temps zéro (T0). Cette possibilité a été abandonné étant donné la volonté de conserver l'aspect temporel associé aux vagues de la pandémie dans la formation des groupes. Il pourrait donc être intéressant de réaliser les regroupements avec ce T0 et de le comparer avec les résultats obtenus en considérant la temporalité de la pandémie.

Ensuite, en ce qui concerne les deux modèles plus classiques basés sur les délais entre les événements plutôt que sur les séquences temporelles, le projet a démontré qu'avec les données de l'étude, il n'y avait pas de différence notable entre les deux. Bien qu'en essence différents (données standardisées versus données non standardisées, classification ascendante versus descendante, matrice de distance Euclidienne versus Manhattan), les

résultats laissent penser qu'il n'est pas systématiquement utile de tester toutes les déclinaisons possibles des différents algorithmes. Une sélection des principales configurations à tester suivi d'un approfondissement des plus prometteuses apparaissait être une approche plus efficiente.

À l'exception du modèle exploitant le *Dynamic time warping*, les modèles retenus avaient tous deux groupes. Malgré qu'un nombre limité de groupe s'avère parfois plus simple à analyser ou interpréter, un nombre un peu plus grand est susceptible de mettre en évidence des tendances ou des caractéristiques plus subtiles ou, encore, des sous-groupes, qui pourraient passer inaperçu au sein d'un groupe plus imposant. C'est, entre autres, ce qui a justifié la présentation des résultats de ce modèle au Chapitre 4. Comparativement aux autres modèles, les cinq groupes captaient et nuançaient mieux la progression temporelle, dont l'importance s'avérait centrale dans le contexte, des événements dans les vagues pandémiques. De toutes les combinaisons testées, ce modèle avait le coefficient de silhouette le plus élevé, suggérant que les regroupements issus de ce dernier étaient à la fois bien séparés et cohérents, ce qui pouvait favoriser la pertinence et la fiabilité des interprétations pouvant être tirées de cette classification.

Outre les éléments présentés précédemment, d'autres éléments intéressants ont été mis en lumière, par exemple, les différences de caractéristiques et de séquences des travailleurs de la santé entre les différents groupes. Ces autres éléments, qui pourront faire l'objet de projets complémentaires, montrent que les techniques non-supervisées, bien qu'elles n'apportent pas toujours de réponses précises, mettent en lumière des pistes intéressantes d'investigation future. Leur utilisation permet donc, entre autres chose, de guider les efforts d'investigation dans le domaine.

6.3 Forces et limites

En termes de forces, les données utilisées pour l'étude ont été collectés au tout début de la pandémie, ce qui en fait une source précieuse d'informations pour un tel projet. La gestion des données est également une force étant donné qu'un travail substantiel a été réalisé afin d'augmenter la taille de l'échantillon à l'étude. De surcroît, l'utilisation de l'apprentissage automatique a permis d'identifier des schémas plus complexes en tenant

compte des événements et de leur temporalité, via une approche « *data-driven* ». En effet, la méthodologie utilisée a permis de considérer, non seulement les délais entre les événements, mais également leur chronologie afin de tenir compte notamment de la vague pandémique dans laquelle l'événement est survenu. Ainsi, le fait de d'abord former les groupes, puis de les caractériser à l'aide de variables qui n'ont pas été utilisées lors de leur constitution initiale a permis de mettre en lumière des éléments potentiellement plus difficiles à repérer avec les méthodes plus classiques. Également, le recours à différents algorithmes a notamment permis d'explorer différentes approches, offrant des perspectives supplémentaires et complémentaires quant résultats. De surcroît, les analyses de sensibilité et de stabilité réalisées constituent une force du projet puisqu'elles ont permis de valider et nuancer les différentes approches et leurs résultats, ce qui octroyait un niveau d'assurance plus élevé quant à leur cohérence et leur robustesse. Elles ont également offert un certain recul face aux algorithmes dans un optique de réponse au sous-objectif d'explorer les différentes combinaisons d'algorithmes et de mesures de dissimilarité pertinentes pour le type de données du projet.

Au chapitre des limites, les données disponibles pouvaient être considérées comme une des limites du projet, notamment en termes de la quantité de données disponibles pour inclure davantage d'individus dans la cohorte qu'en termes d'effort requis en nettoyage et transformation des données. La collaboration avec l'équipe de la BQC19 pour la correction des éléments problématiques a diminué les biais potentiels des données et permis de diminuer la réduction de la taille d'échantillon. De plus, les efforts en nettoyage de données déployées pour compenser les limites ont généré un script réutilisable qui, à l'heure actuelle, a déjà été utilisée pour un projet connexe sur la COVID longue. Aussi, il est possible d'évoquer comme limite le traitement afférent à la date de réinfection. Puisqu'un choix entre deux variables dans le jeu de données a été effectué pour établir celle retenue pour l'étude, il existe effectivement un risque de biais. Toutefois, la méthode utilisée pour l'établir minimise à notre sens les risques étant donné que dans la majorité des cas, les deux dates étaient les mêmes, que le traitement en cas de discordance a été appliquée uniformément pour tous les individus et que les deux dates figuraient déjà dans le jeu de données, ce qui apparaît moins impactant que d'avoir envisagé l'imputation pour une variable de cette importance (ou encore le retrait d'individus supplémentaires). Nous

reconnaissons également que le délai utilisé pour considérer une réinfection subséquente constitue une source de biais potentiel, dont le risque principal est la surévaluation du nombre de réinfection suivant la première réinfection. Considérant l'absence de consensus concernant ce délai dans la communauté scientifique, le délai usuel de la BQC19 a été utilisé sachant que même avec ce court délai de 14 jours, le nombre de patients concernés était relativement petit et, qu'après étude de la distribution des variables des délais interréinfections, il est apparu que le fait d'augmenter le seuil de considération n'avait qu'un impact minime sur le nombre d'individus concernés. Bien que le risque afférent apparaissait relativement faible dans ce contexte, les interprétations et constats se sont limités à ce qui était commun à la cohorte complète, soit les événements jusqu'à la réinfection inclusivement, excluant ainsi les réinfections subséquentes. Cet élément vient également atténuer la durée du suivi qui n'était pas uniforme d'un individu à l'autre et pourrait exercer une influence sur les résultats et leur interprétation, notamment en sous-estimant le nombre d'événement, particulièrement en termes de vaccins et de réinfections.

En somme, pour les données du projet, le *Dynamic time warping* s'est révélé l'algorithme de mesure de dissimilarité le plus pertinent pour répondre à l'objet de l'étude parmi les différentes combinaisons testées. Il offrait plus de nuance dans la composition de groupes comparativement aux algorithmes plus classiques. Étonnement, l'*Optimal matching* s'est révélé l'algorithme le moins performant bien que le contexte permettait a priori de croire le contraire puisqu'il s'agissait de séquences d'événements. La caractérisation des groupes a notamment permis de confirmer le rôle de la vaccination dans les séquences. Ces résultats étaient cohérents avec plusieurs éléments de la littérature et de l'état des connaissances actuelles. Il s'agit d'une perspective intéressante en termes d'apprentissage à tirer des politiques en réponse à la pandémie de COVID-19 et, par le fait même, ouvre une perspective de préparation aux futures pandémies potentielles, notamment en soulevant des questions qui auraient pu rester dans l'ombre dans le cas d'utilisation de méthodes statistiques classiques.

Sur le plan des perspectives de recherche, il pourrait être intéressant de bénéficier des nouvelles versions des données publiées par la BQC19, notamment dans un espoir

d'élargir le nombre de patients répondant aux critères de la cohorte. Le cas échéant, la méthodologie pourrait alors être reproduite sur un plus grand nombre de participants. De plus, les groupes résultants du projet pourraient être caractérisés à l'aide de données biologiques, lorsqu'elles seront disponibles pour la cohorte, afin de déterminer si ces dernières pourraient fournir des explications supplémentaires des différents profils de réinfection des individus. Le cas échéant, l'utilisation des forêts aléatoires (« *random forest* ») est envisagée pour déterminer quelles nouvelles variables génomiques auraient le plus grand pouvoir explicatif. De même, le groupe d'appartenance des participants de l'étude pourrait être utilisée comme variables dans d'autres projets, comme le projet apparenté sur le syndrome post-COVID (COVID longue), afin de déterminer si les profils de réinfection ou la séquence d'événements ayant mené à l'immunité influence les manifestations ou les résultats des autres études.

Chapitre 7 Conclusion

La pandémie de COVID-19 a mis en lumière le rôle crucial de la santé publique dans les systèmes de santé. Face à cette crise sanitaire mondiale sans précédent, l'importance de la prise de décision rapide fondée sur des données probantes fût particulièrement mise de l'avant. Dans le contexte actuel où les mégadonnées (le « *big data* ») et l'intelligence artificielle gagnent en influence dans une multitude de sphères de la société, la science des données apparaît comme l'un des piliers incontournables des systèmes de santé de demain. Dans ce contexte, les initiatives comme celle de la Biobanque québécoise de la COVID19, en collectant et démocratisant l'accès aux données, permettent non seulement d'étudier les maladies et la santé de population, mais également de supporter la prise de décision en matière de gestion et de politiques de santé afférentes pour le système de la santé du Québec. Elles permettent conséquemment aux systèmes d'apprendre afin de se transformer et d'évoluer au bénéfice de la santé des populations.

Au terme du projet, quatre modèles ont été sélectionnés sur la base de coefficient de Silhouette moyen parmi la multitude de combinaisons de paramètres différents testés lors des analyses de regroupement. Les groupes issus de ces modèles ont été caractérisés en termes de variables sociodémographiques, cliniques, de délais interévénements et de séquences. L'algorithme de plus pertinent aux données et au contexte de l'étude a été identifié. Il a regroupé les patients de la cohorte en cinq groupes à partir des seules séquences temporelles des événements d'intérêts. L'étude des caractéristiques et des séquences temporelles des groupes a permis de mettre en lumière le rôle des politiques de vaccination dans les profils d'infection et de réinfection.

Quoi qu'il en soit, la valorisation des données de la BQC19 combinée aux techniques d'apprentissage automatique constitue, comme le montre le projet, une source précieuse d'informations pour la recherche dans le domaine de la santé des populations. La valorisation des données, via l'apprentissage non supervisé, a notamment soulevé divers questionnements qui serviront de point de départ à des investigations futures. Au-delà de cet apport, elle constitue également un levier stratégique pour optimiser la gestion,

améliorer la performance et orienter l'évolution des systèmes de santé. L'exploitation de données probantes favorise la prise de décision plus éclairée, mais constitue également une approche prometteuse afin de répondre aux défis croissants des systèmes de santé. Dans cette perspective, l'intégration accrue de la science des données dans les pratiques quotidiennes des établissements de santé et la collaboration entre chercheurs, cliniciens et gestionnaires représentent un champ d'investigation et d’action à exploiter.

Bibliographie

Abbott, Andrew et John Forrest (1986). « Optimal Matching Methods for Historical Sequences », *The Journal of Interdisciplinary History*, vol. 16, no 3, p. 471-494.

Abbott, Andrew et Alexandra Hrycak (1990). « Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers », *American Journal of Sociology*, vol. 96, no 1, p. 144-185.

Agence ontarienne de protection et de promotion de la santé (2023). *Variants préoccupants de la COVID-19*. Récupéré le 22 septembre 2024

Ait Abdelouahid, Rachida, Olivier Debauche, Saïd Mahmoudi et Abdelaziz Marzak (2023). « Literature Review: Clinical Data Interoperability Models », *Information*, vol. 14, no 7.

Alanazi, Abdullah (2022). « Using machine learning for healthcare challenges and opportunities », *Informatics in Medicine Unlocked*, vol. 30.

Aleem, A., A. B. Akbar Samad et S. Vaqar (2024). « Emerging Variants of SARS-CoV-2 and Novel Therapeutics Against Coronavirus (COVID-19) », dans *StatPearls*, Treasure Island (FL), StatPearls Publishing Copyright © 2024, StatPearls Publishing LLC.

Arabi, M., Y. Al-Najjar, O. Sharma, I. Kamal, A. Javed, H. S. Gohil*, et al.* (2023). « Role of previous infection with SARS-CoV-2 in protecting against omicron reinfections and severe complications of COVID-19 compared to pre-omicron variants: a systematic review », *BMC Infect Dis*, vol. 23, no 1, p. 432.

Asamoah-Boaheng, M., B. Grunau, S. Haig, M. E. Karim, T. Kirkham, P. M. Lavoie*, et al.* (2023). « Eleven-month SARS-CoV-2 binding antibody decay, and associated factors, among mRNA vaccinees: implications for booster vaccination », *Access Microbiol*, vol. 5, no 11.

Beaulieu, Martin, Jacques Roy, Claudia Rebolledo et Sylvain Landry (2021). *Gestion des équipements de protection dans le réseau québécois de la santé : chronologie des évènements, constats et recommandations*,

Bhatia, AshishSingh et Yu-Wei Chiu (2017). *Machine Learning with R Cookbook*, 2e éd., Packt Publishing Ltd.

Biobanque québécoise de la COVID-19 (2022a). *Matériel disponible*. Récupéré le 15 juillet 2024 https://www.quebeccovidbiobank.ca/materiel-disponible

Biobanque québécoise de la COVID-19 (2022b). *Présentation de la BQC19*. Récupéré le 10 décembre 2023 2023 de https://www.quebeccovidbiobank.ca/a-propos

Biobanque québécoise de la Covid-19 (s.d.). *BQC19 - Design étude*. Récupéré le 19 juillet 2024 https://www.bqc19.ca/fr/design-etude

Boaventura, Viviane S., Thiago Cerqueira-Silva et Manoel Barral-Netto (2023). « The benefit of vaccination after previous SARS-CoV-2 infection in the omicron era », *The Lancet Infectious Diseases*, vol. 23, no 5, p. 511-512.

Bobrovitz, N., H. Ware, X. Ma, Z. Li, R. Hosseini, C. Cao*, et al.* (2023). « Protective effectiveness of previous SARS-CoV-2 infection and hybrid immunity against the omicron variant and severe disease: a systematic review and meta-regression », *Lancet Infect Dis*, vol. 23, no 5, p. 556-567.

Bobrovitz, Niklas, Harriet Ware, Xiaomeng Ma, Zihan Li, Reza Hosseini, Christian Cao*, et al.* (2023). « Protective effectiveness of previous SARS-CoV-2 infection and hybrid immunity against the omicron variant and severe disease: a systematic review and meta-regression », *The Lancet Infectious Diseases*, vol. 23, no 5, p. 556-567.

Bonny, V., A. Maillard, C. Mousseaux, L. Plaçais et Q. Richier (2020). « COVID-19 : physiopathologie d'une maladie à plusieurs visages », *La Revue de Médecine Interne*, vol. 41, no 6, p. 375-389.

Bush, Larry M. (2022). *Développement d'une infection - Infections*, Manuel Merck. Récupéré le 18 juillet 2024 https://www.merckmanuals.com/fr-ca/accueil/infections/biologie-des-maladies-infectieuses/d%C3%A9veloppement-d%E2%80%99une-infection

Canadian Institute for Health Information (2021a). *Overview: COVID-19's impact on health care systems*. Récupéré le 21 septembre 2024 https://www.cihi.ca/en/covid-19-resources/impact-of-covid-19-on-canadas-health-care-systems/the-big-picture

Canadian Institute for Health Information (2021b). *Overview: Impacts of COVID-19 on health care providers*. Récupéré le 21 septembre 2024 https://www.cihi.ca/en/health-workforce-in-canada-in-focus-including-nurses-and-physicians/overview-impacts-of-covid-19-on

Canoui, E. et O. Launay (2019). « History and principles of vaccination », *Rev Mal Respir*, vol. 36, no 1, p. 74-81.

Carazo, S., D. M. Skowronski, M. Brisson, S. Barkati, C. Sauvageau, N. Brousseau*, et al.* (2023). « Protection against omicron (B.1.1.529) BA.2 reinfection conferred by primary omicron BA.1 or pre-omicron SARS-CoV-2 infection among health-care workers with and without mRNA vaccination: a test-negative case-control study », *Lancet Infect Dis*, vol. 23, no 1, p. 45-55.

Cascella, M., M. Rajnik, A. Aleem, S. C. Dulebohn et R. Di Napoli (2024). « Features, Evaluation, and Treatment of Coronavirus (COVID-19) », dans *StatPearls*, Treasure Island (FL), StatPearls Publishing LLC.

Centers for Disease Control and Prevention (2024). *About Reinfection - COVID-19*. Récupéré le 17 juillet 2024 https://www.cdc.gov/covid/about/reinfection.html?CDC_AAref_Val=https://www.cdc.gov/coronavirus/2019-nc

Chemaitelly, Hiam, Houssein H. Ayoub, Patrick Tang, Hadi M. Yassine, Asmaa A. Al Thani, Mohammad R. Hasan*, et al.* (2024). « Addressing bias in the definition of SARS-CoV-2 reinfection: implications for underestimation », *Frontiers in Medicine*, vol. 11.

Chen, Y., W. Zhu, X. Han, M. Chen, X. Li, H. Huang*, et al.* (2024). « How does the SARS-CoV-2 reinfection rate change over time? The global evidence from systematic review and meta-analysis », *BMC Infect Dis*, vol. 24, no 1, p. 339.

Comité sur l'immunisation du Québec (2020). *Stratégie de vaccination contre la COVID-19 : report de la 2e dose en contexte de pénurie*, Montréal Québec, Institut national de santé publique du Québec, 7 p. Récupéré de https://www.inspq.qc.ca/publications/3098-strategie-vaccination-2e-dose-covid

Comité sur l'immunisation du Québec (2021). *Avis préliminaire sur les groupes prioritaires pour la vaccination contre la COVID-19 au Québec*, Montréal Québec, Institut national de santé publique du Québec, 76 p. Récupéré de https://www.inspq.qc.ca/publications/3085-groupes-prioritaires-vaccination-covid

Comité sur l'immunisation du Québec, Nicholas Brousseau, Marilou Kiely et Sara Carazo (2024). *Vaccination contre la COVID-19 : Recommandations pour l'automne 2024 : avis scientifique intérimaire*, Montréal Québec, Institut national de santé publique du Québec.

Comité sur l'immunisation du Québec, Philippe De Wals, Rodica Gilca, Gaston De Serres, Nicholas Brousseau, Caroline Quach-Thanh*, et al.* (2022). *Stratégie vaccinale contre la COVID-19 à préconiser au Québec en 2022 et pertinence d'une 2e dose de rappel pour certains groupes vulnérables*, Montréal Québec, Institut national de santé publique du Québec, 22 p. Récupéré de https://www.inspq.qc.ca/publications/3207-2e-dose-rappel-covid

Crits-Christoph, Alexander, Joshua I. Levy, Jonathan E. Pekar, Stephen A. Goldstein, Reema Singh, Zach Hensel*, et al.* (2024). « Genetic tracing of market wildlife and viruses at the epicenter of the COVID-19 pandemic », *Cell*, vol. 187, no 19, p. 5468-5482.e5411.

Diani, S., E. Leonardi, A. Cavezzi, S. Ferrari, O. Iacono, A. Limoli*, et al.* (2022). « SARS-CoV-2-The Role of Natural Immunity: A Narrative Review », *J Clin Med*, vol. 11, no 21.

Dutta, Abhishek (2022). « COVID-19 waves: variant dynamics and control », *Scientific Reports*, vol. 12, no 1, p. 9332.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander et Xiaowei Xu (1996). « A density-based algorithm for discovering clusters in large spatial databases with noise », communication présentée au *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon.

European Centre for Disease Prevention and Control (2020). *Reinfection with SARS-CoV: considerations for public health response: ECDC*. Récupéré de https://www.ecdc.europa.eu/sites/default/files/documents/Re-infection-and-viral-shedding-threat-assessment-brief.pdf

Fox, J. et S. Weisberg (2019). *An R Companion to Applied Regression*, 3e éd., Thousand Oaks CA. Récupéré de https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas S. Müller et Matthias Studer (2011). « Analyzing and Visualizing State Sequences in R with TraMineR », *Journal of Statistical Software*, vol. 40, no 4, p. 1 - 37.

Giorgino, Toni (2009). « Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package », *Journal of Statistical Software*, vol. 31, no 7, p. 1 - 24.

Gostin Lawrence, O. et K. Gronvall Gigi (2023). « The Origins of Covid-19 — Why It Matters (and Why It Doesn't) », *New England Journal of Medicine*, vol. 388, no 25, p. 2305-2308.

Gouvernement du Canada (2024). *Syndrome post-COVID-19 (COVID longue)*. Récupéré le 18 juillet 2024 https://www.canada.ca/fr/sante-publique/services/maladies/2019-nouveau-coronavirus/symptomes/syndrome-post-covid-19.html

Gouvernement du Québec (2023). *Comprendre la vaccination*. 16 juillet 2024 de https://www.quebec.ca/sante/conseils-et-prevention/vaccination/comprendre-la-vaccination

Gouvernement du Québec (2024). *Vaccination contre la COVID-19*. Récupéré le 22 septembre 2024 https://www.quebec.ca/sante/conseils-et-prevention/vaccination/vaccin-contre-la-covid-19#c233194

Gralinski, L. E. et V. D. Menachery (2020). « Return of the Coronavirus: 2019-nCoV », *Viruses*, vol. 12, no 2.

Groupe de travail sur l'immunité face à la COVID-19 (2021). *Determining the impact of hybrid immunity on the evolving landscape of host responses to SARS-CoV-2 in the Biobanque Québécoise de la COVID-19 (BQC19)*. Récupéré le 24 avril 2023 https://www.covid19immunitytaskforce.ca/fr/recherche-du-groupe-de-travail/immunite-hybride/

Groupe de travail sur l'immunité face à la COVID-19 (2024). *Foire aux questions*. Récupéré le 30 novembre 2024 https://www.covid19immunitytaskforce.ca/fr/faq/

Guimaraes, L. E., B. Baker, C. Perricone et Y. Shoenfeld (2015). « Vaccines, adjuvants and autoimmunity », *Pharmacol Res*, vol. 100, p. 190-209.

Hastie, Trevor J., Robert John Tibshirani, Jerome H. Friedman, Robert Tibshirani et Jerome Friedman (2009). *The elements of statistical learning : data mining, inference, and prediction*, Second edition[e] éd., New York, NY, Springer New York : Springer e-books : Imprint : Springer : Springer e-books. Récupéré de https://doi.org/10.1007/b94608

He, Zhenyu, Lili Ren, Juntao Yang, Li Guo, Luzhao Feng, Chao Ma*, et al.* (2021). « Seroprevalence and humoral immune durability of anti-SARS-CoV-2 antibodies in Wuhan, China: a longitudinal, population-level, cross-sectional study », *The Lancet*, vol. 397, no 10279, p. 1075-1084.

Henderson, Donald A. (2011). « The eradication of smallpox – An overview of the past, present, and future », *Vaccine*, vol. 29, p. D7-D9.

Hennig, Christian (2023). *fpc : Flexible Procedures for Clustering. R Package*, version 2.2-11, https://cran.r-project.org/web/packages/fpc/index.html

Hirano, Shoji, Xiaoguang Sun et Shusaku Tsumoto (2004). « Comparison of clustering methods for clinical databases », *Information Sciences*, vol. 159, no 3, p. 155-165.

Hu, B., H. Guo, P. Zhou et Z. L. Shi (2021). « Characteristics of SARS-CoV-2 and COVID-19 », *Nat Rev Microbiol*, vol. 19, no 3, p. 141-154.

Huang, C., Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu*, et al.* (2020). « Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China », *Lancet*, vol. 395, no 10223, p. 497-506.

Institut national de santé publique du Québec (2023). *Les variants du SRAS-CoV-2*. Récupéré le 22 septembre 2024 https://www.inspq.qc.ca/covid-19/labo/variants

Institut national de santé publique du Québec (2024). *Méthodologie des données COVID-19*. Récupéré le 22 septembre 2024 https://www.inspq.qc.ca/covid-19/donnees/methodologie

Institut national de santé publique du Québec (s.d.). *COVID-19 : virus et transmission*. Récupéré le 16 juillet 2024 https://www.inspq.qc.ca/sante-voyage/guide/immunisation/covid-19/virus-covid-19-sa-transmission

Jacobsen, H., I. Sitaras, M. Katzmarzyk, V. Cobos Jiménez, R. Naughton, M. M. Higdon*, et al.* (2023). « Systematic review and meta-analysis of the factors affecting waning of post-vaccination neutralizing antibody responses against SARS-CoV-2 », *NPJ Vaccines*, vol. 8, no 1, p. 159.

Janssenswillen, Gert, Benoît Depaire, Marijke Swennen, Mieke Jans et Koen Vanhoof (2019). « bupaR: Enabling reproducible business process analysis », *Knowledge-Based Systems*, vol. 163, p. 927-930.

Karia, R., I. Gupta, H. Khandait, A. Yadav et A. Yadav (2020). « COVID-19 and its Modes of Transmission », *SN Compr Clin Med*, vol. 2, no 10, p. 1798-1801.

Kassambara, Alboukadel (2023). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, version 0.7.2, https://cran.r-project.org/web/packages/rstatix/index.html

Kassambara, Alboukadel et Fabian Mundt (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package*, version 1.0.7, https://cran.r-project.org/web/packages/factoextra/index.html

Keeling, Matt J. (2023). « Patterns of reported infection and reinfection of SARS-CoV-2 in England », *Journal of Theoretical Biology*, vol. 556, p. 111299.

Koirala, A., Y. J. Joo, A. Khatami, C. Chiu et P. N. Britton (2020). « Vaccines for COVID-19: The current state of play », *Paediatr Respir Rev*, vol. 35, p. 43-49.

Kolde, Raivo (2019). *pheatmap: Pretty Heatmaps. R package*, version 1.0.12, 10.32614/CRAN.package.pheatmap

Lapuente, D., T. H. Winkler et M. Tenbusch (2024). « B-cell and antibody responses to SARS-CoV-2: infection, vaccination, and hybrid immunity », *Cell Mol Immunol*, vol. 21, no 2, p. 144-158.

Lee, In et Yong Jae Shin (2020). « Machine learning for enterprises: Applications, algorithm selection, and challenges », Business Horizons, vol. 63, no 2, p. 157-170.

Livieratos, A., C. Gogos et K. Akinosoglou (2024). « Impact of Prior COVID-19 Immunization and/or Prior Infection on Immune Responses and Clinical Outcomes », *Viruses*, vol. 16, no 5.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert et K. Hornik (2023). *cluster : Cluster Analysis Basics and Extensions. R package*, version 2.1.6, https://CRAN.R-project.org/package=cluster.

Mahesh, Batta (2020). « Machine Learning Algorithms - A Review », *International Journal of Science and Research (IJSR)*, vol. 9, no 1, p. 381-386.

Maragakis, Lisa (2021). *Coronavirus Second Wave, Third Wave and Beyond: What Causes a COVID Surge*, Johns Hopkins Medicine. Récupéré le 22 septembre 2024 https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/first-and-second-waves-of-coronavirus

McGill University (2020). *Le FRQS constitue une biobanque panquébécoise de la COVID-19*. Récupéré le 14 août 2023 https://www.mcgill.ca/newsroom/fr/channels/news/le-frqs-constitue-une-biobanque-panquebecoise-de-la-covid-19-321346

Meyer, David. et Christian. Buchta (2022). *proxy: Distance and Similarity Measures. R Package*, version 0.4-27, https://cran.r-project.org/web/packages/proxy/index.html

Ministère de la Santé et des Services sociaux (2018). *Fonctionnement du système immunitaire - Immunologie de la vaccination - Professionnels de la santé*. Récupéré le 18 juillet 2024 https://www.msss.gouv.qc.ca/professionnels/vaccination/piq-immunologie-de-la-vaccination/fonctionnement-du-systeme-immunitaire/

Ministère de la Santé et des Services sociaux (2020). *Définitions - Immunologie de la vaccination - Professionnels de la santé*. 16 juillet 2024 de https://www.quebec.ca/sante/conseils-et-prevention/vaccination/comprendre-la-vaccination

Misra, A. et E. S. Theel (2022). « Immunity to SARS-CoV-2: What Do We Know and Should We Be Testing for It? », *J Clin Microbiol*, vol. 60, no 6, p. e0048221.

Moghadas, S. M., T. N. Vilches, K. Zhang, C. R. Wells, A. Shoukat, B. H. Singer*, et al.* (2021). « The Impact of Vaccination on Coronavirus Disease 2019 (COVID-19) Outbreaks in the United States », *Clin Infect Dis*, vol. 73, no 12, p. 2257-2264.

Munoz-Gama, J., N. Martin, C. Fernandez-Llatas, O. A. Johnson, M. Sepúlveda, E. Helm*, et al.* (2022). « Process mining for healthcare: Characteristics and challenges », *J Biomed Inform*, vol. 127, p. 103994.

National Cancer Institute (s.d.). *Definition of infection - NCI Dictionary of Cancer Terms*. Récupéré le 18 juillet 2024 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/infection

National Institute of Allergy And Infectious Diseases (s.d.). *Coronaviruses*. Récupéré le 16 juillet 2024 https://www.niaid.nih.gov/diseases-conditions/coronaviruses

National Library of Medecine (s.d.). *Infection (Concept Id: C3714514) - MedGen - NCBI*. Récupéré le 18 juillet 2024 https://www.ncbi.nlm.nih.gov/medgen/811352#Definition

Nayyar, Anand, Lata Gadhavi et Noor Zaman (2021). « Machine learning in healthcare: review, opportunities and challenges », dans Krishna Kant Singh, Mohamed Elhoseny, Akansha Singh et Ahmed A. Elngar (dir.), *Machine Learning and the Internet of Medical Things in Healthcare*, Academic Press, p. 23-45.

Ozaydin, Bunyamin, Eta S. Berner et James J. Cimino (2021). « Appropriate use of machine learning in healthcare », Intelligence-Based Medicine, vol. 5, p. 100041.

Pilz, S., V. Theiler-Schwetz, C. Trummer, R. Krause et J. P. A. Ioannidis (2022). « SARS-CoV-2 reinfections: Overview of efficacy and duration of natural and hybrid immunity », *Environ Res*, vol. 209, p. 112911.

Plotkin, Stanley (2014). « History of vaccination », *Proceedings of the National Academy of Sciences*, vol. 111, no 34, p. 12283-12287.

Posit team (2024). *RStudio: Integrated Development Environment for R*, PBC, Boston, MA, Posit Software, http://www.posit.co/.

Pyle, Dorian (1999). *Data Preparation for Data Mining*, Morgan Kaufmann Publishers Inc.

R Core Team, . (2023). *R : A Language and Environment for Statistical Computing*, Vienna, Austria, R Foundation for Statistical Computing, https://www.R-project.org/

Raab M (2022). *ggseqplot: Render Sequence Plots using 'ggplot2. R Package*, https://maraab23.github.io/ggseqplot/.

Rambaut, A., E. C. Holmes, A. O'Toole, V. Hill, J. T. McCrone, C. Ruis*, et al.* (2020). « A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology », *Nat Microbiol*, vol. 5, no 11, p. 1403-1407.

Rebala, Gopinath, Ajay Ravi et Sanjay Churiwala (2019). *An introduction to machine learning*, Cham, Springer. Récupéré de https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2120384

Rodriguez, M. Z., C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. D. F. Costa*, et al.* (2019). « Clustering algorithms: A comparative approach », *PLOS ONE*, vol. 14, no 1, p. e0210236.

Rodriguez Velásquez, S., L. E. Biru, S. M. Hakiza, M. Al-Gobari, I. Triulzi, J. Dalal*, et al.* (2024). « Long-term levels of protection of different types of immunity against the Omicron variant: a rapid literature review », *Swiss Med Wkly*, vol. 154, p. 3732.

Röltgen, K., A. E. Powell, O. F. Wirz, B. A. Stevens, C. A. Hogan, J. Najeeb*, et al.* (2020). « Defining the features and duration of antibody responses to SARS-CoV-2 infection associated with disease severity and outcome », *Sci Immunol*, vol. 5, no 54.

Rousseeuw, Peter J. (1987). « Silhouettes: A graphical aid to the interpretation and validation of cluster analysis », *Journal of Computational and Applied Mathematics*, vol. 20, p. 53-65.

Sanchez-de Prada, L., A. M. Martinez-Garcia, B. Gonzalez-Fernandez, J. Gutierrez-Ballesteros, S. Rojo-Rello, S. Garcinuno-Perez*, et al.* (2024). « Impact on the time elapsed since SARS-CoV-2 infection, vaccination history, and number of doses, on protection against reinfection », *Sci Rep*, vol. 14, no 1, p. 353.

Sander, Jörg, Martin Ester, Hans-Peter Kriegel et Xiaowei Xu (1998). « Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications », *Data Mining and Knowledge Discovery*, vol. 2, no 2, p. 169-194.

Santé Canada (2003). « Canadian Guidelines for Body Weight Classification in Adults ». https://www.canada.ca/en/health-canada/services/food-nutrition/healthy-eating/healthy-weights/canadian-guidelines-body-weight-classification-adults/body-mass-index-nomogram.html

Sarda-Espinosa, A. (2023). *dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance. R package* version 5.5.12, https://CRAN.R-project.org/package=dtwclust

Sardá-Espinosa, Alexis (2017). « Comparing time-series clustering algorithms in r using the dtwclust package », *R package vignette*, vol. 12, p. 41.

Scientifique en chef du Québec (2020). *Création de la Biobanque québécoise de COVID*. Récupéré le 19 juillet 2024 https://www.scientifique-en-chef.gouv.qc.ca/creation-de-la-biobanque-quebecoise-de-covid/

Scrucca, Luca, Chris Fraley, Thomas Murphy et Raftery E (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*,

Selker, Ravi, Jonathon Love, Damian Dropmann, Victor Moreno et Maurizio Agosti (2024). *jmv: The 'jamovi' Analyses. R Package*, version 2.5.6, https://cran.r-project.org/web/packages/jmv/index.html

Statistique Canada (2022). *Classification des catégories d'âge par tranches de cinq ans, variante de regroupement, avec détails pour 18 et plus*. Récupéré le 14 novembre 2023 https://www23.statcan.gc.ca/imdb/p3VD_f.pl?Function=getVD&TVD=491082

Strassburg, Marc A. (1982). « The global eradication of smallpox », *American Journal of Infection Control*, vol. 10, no 2, p. 53-59.

Studer, Matthias (2013). « WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R », *LIVES Working papers*, vol. 24.

Studer, Matthias et Gilbert Ritschard (2015). « What Matters in Differences Between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures », *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 179, no 2, p. 481-511.

Studer, Matthias et Gilbert Ritschard (2016). « What Matters in Differences Between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures », *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 179, no 2, p. 481-511.

Stulpin, Caitlyn (2023). « Hybrid COVID-19 immunity offers higher protection vs. previous infection alone », *Infectious Disease News*, vol. 36, no 3, p. 44.

Tremblay, Karine, Simon Rousseau, Ma'n H. Zawati, Daniel Auld, Michaël Chassé, Daniel Coderre*, et al.* (2021). « The Biobanque québécoise de la COVID-19 (BQC19)—A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories », *PLOS ONE*, vol. 16, no 5, p. e0245031.

Tufféry, Stéphane (2010). *Data mining et statistique décisionnelle : l'intelligence des données*, 3e. éd. actualisée et augm^{e} éd., Paris, Éditions Technip.

Walsh, I., D. Fishman, D. Garcia-Gasulla, T. Titma, G. Pollastri, Elixir Machine Learning Focus Group, et al. (2021). « DOME: recommendations for supervised machine learning validation in biology », Nat Methods, vol. 18, no 10, p. 1122-1127.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François*, et al.* (2019). « Welcome to the Tidyverse », *Journal of Open Source Software*, vol. 4, no 43.

World Health Organization (2020a). *Chronologie de l'action de l'OMS face à la COVID-19*. Récupéré le 14 août 2023 https://www.who.int/fr/news/item/29-06-2020-covidtimeline

World Health Organization (2020b). *Naming the coronavirus disease (COVID-19) and the virus that causes it*. Récupéré le 16 juillet 2024 https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it

World Health Organization (2020c). *Situation report - Novel Coronavirus (2019-nCoV) - #01 - 21 January 2020*. Récupéré de https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4

World Health Organization (2020d). *Situation report - Novel Coronavirus (2019-nCoV) - #07 - 27 January 2020*, no 7. Récupéré de https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200127-sitrep-7-2019--ncov.pdf?sfvrsn=98ef79f5_2

World Health Organization (2020e). *Situation report - Novel Coronavirus (2019-nCoV) - #12 - 1 February 2020*. Récupéré de https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200201-sitrep-12-ncov.pdf?sfvrsn=273c5d35_2

World Health Organization (2020f). *Situation report - Novel Coronavirus (2019-nCoV) - #26 - 15 February 2020*. Récupéré de https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200215-sitrep-26-covid-19.pdf?sfvrsn=a4cc6787_2

World Health Organization (2020g). *Situation report - Novel Coronavirus (2019-nCoV) - #51 - 11 March 2020*. Récupéré de https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57_10

World Health Organization (2024). *WHO COVID-19 dashboard - Data reported on 28 april 2024*. Récupéré le 17 juillet 2024 https://data.who.int/dashboards/covid19/

World Health Organization (s.d.-a). *Coronavirus*. Récupéré le 16 juillet 2024 https://www.who.int/fr/health-topics/coronavirus/

World Health Organization (s.d.-b). *Vaccins et vaccination - Impact*. 16 juillet 2024 de https://www.who.int/fr/health-topics/vaccines-and-immunization#tab=tab_2

World Health Organization (s.d.-c). *Vaccins et vaccination - Vue d'ensemble*. 16 juillet 2024 de https://www.who.int/fr/health-topics/vaccines-and-immunization#tab=tab_1

World Health Organization et Igor Kryuchkov (2022). *Post COVID-19 condition (Long COVID)*. Récupéré le 9 septembre 2024 https://www.who.int/europe/news-room/fact-sheets/item/post-covid-19-condition

Worobey, M., J. I. Levy, L. Malpica Serrano, A. Crits-Christoph, J. E. Pekar, S. A. Goldstein*, et al.* (2022). « The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic », *Science*, vol. 377, no 6609, p. 951-959.

Wynants, Laure, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, et al. (2020). « Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal », BMJ, vol. 369, p. m1328.

Yahav, D., D. Yelin, I. Eckerle, C. S. Eberhardt, J. Wang, B. Cao*, et al.* (2021). « Definitions for coronavirus disease 2019 reinfection, relapse and PCR re-positivity », *Clin Microbiol Infect*, vol. 27, no 3, p. 315-318.

Zhu, N., D. Zhang, W. Wang, X. Li, B. Yang, J. Song*, et al.* (2020). « A Novel Coronavirus from Patients with Pneumonia in China, 2019 », *N Engl J Med*, vol. 382, no 8, p. 727-733.

Annexes

Annexe 1 Statistiques descriptives de la population

	Homme (N = 2 754)	Femme (N = 3 517)	Population (N = 6 271)
Âge (année)			
Moyenne (écart-type)	56,0 (20,7)	50,3 (20,6)	52,8 (20,8)
Médiane [min, max]	58,2 [0, 98,7]	48,3 [0, 104]	52,9 [0, 104]
Classe d'âge			
Moins de 1 an	38 (1,4%)	37 (1,1%)	75 (1,2%)
1 à 17 ans	73 (2,7%)	94 (2,7%)	167 (2,7%)
18 à 34 ans	350 (12,7%)	742 (21,1%)	1092 (17,4%)
35 à 44 ans	291 (10,6%)	592 (16,8%)	883 (14,1%)
45 à 64 ans	995 (36,1%)	1 189 (33,8%)	2 184 (34,8%)
65 ans et plus	1 007 (36,6%)	863 (24,5%)	1 870 (29,8%)
Statut tabagique			
Non-fumeur	1 475 (53,6%)	2 219 (63,1%)	3 694 (58,9%)
Fumeur	214 (7,8%)	220 (6,3%)	434 (6,9%)
Ex-fumeur	546 (19,8%)	566 (16,1%)	1 112 (17,7%)
Tabagisme passif	16 (0,6%)	15 (0,4%)	31 (0,5%)
Inconnu	503 (18,3%)	497 (14,1%)	1 000 (15,9%)
Cigarette électronique			
Non	2 297 (83,4%)	3 031 (86,2%)	5 328 (85,0%)
Oui	51 (1,9%)	69 (2,0%)	120 (1,9%)
Inconnu	406 (14,7%)	417 (11,9%)	823 (13,1%)
Usage de drogues			
Non	2 182 (79,2%)	2 975 (84,6%)	5 157 (82,2%)
Oui	205 (7,4%)	151 (4,3%)	356 (5,7%)
Inconnu	367 (13,3%)	391 (11,1%)	758 (12,1%)
Type de participant			
Cohorte maladie sévère	1 889 (68,6%)	1 579 (44,9%)	3 468 (55,3%)
Cohorte maladie peu sévère / asymptomatique	865 (31,4%)	1 938 (55,1%)	2 803 (44,7%)
Sévérité de la maladie			
Léger	1 061 (38,5%)	2 097 (59,6%)	3 158 (50,4%)
Modéré	615 (22,3%)	532 (15,1%)	1 147 (18,3%)
Sévère	332 (12,1%)	168 (4,8%)	500 (8,0%)
Décès	79 (2,9%)	51 (1,5%)	130 (2,1%)
Inconnu	667 (24,2%)	669 (19,0%)	1 336 (21,3%)
Travailleur de la santé			
Non	2 064 (74,9%)	1 855 (52,7%)	3 919 (62,5%)
Oui	310 (11,3%)	1 277 (36,3%)	1 587 (25,3%)
Inconnu	380 (13,8%)	385 (10,9%)	765 (12,2%)
Travailleur de laboratoire			
Non	2 347 (85,2%)	3 088 (87,8%)	5 435 (86,7%)
Oui	15 (0,5%)	30 (0,9%)	45 (0,7%)
Inconnu	392 (14,2%)	399 (11,3%)	791 (12,6%)
Statut vaccinal (selon l'OMS)			
Non adéquatement vacciné	228 (8,3%)	171 (4,9%)	399 (6,4%)
Adéquatement vacciné	1 276 (46,3%)	2 114 (60,1%)	3 390 (54,1%)
Inconnu	1 250 (45,4%)	1 232 (35,0%)	2 482 (39,6%)
Nombre de doses de vaccin (sortie étude)			
Moyenne (écart-type)	2,57 (0,965)	2,69 (0,864)	2,64 (0,905)
Médiane [min, max]	3,00 [1,00, 6,00]	3,00 [1,00, 5,00]	3,00 [1,00, 6,00]
Inconnu	1 498 (54,4%)	1 418 (40,3%)	2 916 (46,5%)
Nombre de doses de vaccin lors de l'infection primaire			
Moyenne (écart-type)	0,558 (1,08)	1,02 (1,31)	0,825 (1,24)

	Homme (N = 2 754)	Femme (N = 3 517)	Population (N = 6 271)
Médiane [min, max]	0 [0, 3,00]	0 [0, 3,00]	0 [0, 3,00]
Inconnu	539 (19,6%)	522 (14,8%)	1 061 (16,9%)
Nombre de doses de vaccin lors de la première réinfection			
Moyenne (écart-type)	1,99 (1,10)	2,41 (0,856)	2,29 (0,953)
Médiane [min, max]	2,00 [0, 3,00]	3,00 [0, 3,00]	3,00 [0, 3,00]
Inconnu	2 651 (96,3%)	3 271 (93,0%)	5 922 (94,4%)
Nombre de doses de vaccin lors de la deuxième réinfection			
Moyenne (écart-type)	2,38 (0,744)	2,29 (0,908)	2,31 (0,859)
Médiane [min, max]	2,50 [1,00, 3,00]	3,00 [0, 3,00]	3,00 [0, 3,00]
Inconnu	2 746 (99,7%)	3 493 (99,3%)	6 239 (99,5%)
Nombre de doses de vaccin lors de la troisième réinfection			
Moyenne (écart-type)	--	2,50 (0,548)	2,50 (0,548)
Médiane [min, max]	--	2,50 [2,00, 3,00]	2,50 [2,00, 3,00]
Inconnu	2 754 (100%)	3 511 (99,8%)	6 265 (99,9%)
Nombre de réinfections			
Moyenne (écart-type)	1,08 (0,269)	1,12 (0,396)	1,11 (0,363)
Médiane [min, max]	1,00 [1,00, 2,00]	1,00 [1,00, 3,00]	1,00 [1,00, 3,00]
Inconnu	2 651 (96,3%)	3 271 (93,0%)	5 922 (94,4%)

Note : Un individu a été retiré en raison de l'intégrité des données (voir section 3.2.1.3)

Annexe 2 BQC19 - Liste complète des variables du jeu de données principal

Adapté du site web de la Biobanque québécoise de la Covid-19 (s.d.).

Cohorte Maladie sévère (hospitalisée)	Hospitalisation							Suivi post-congé hospitalier						Sortie
Cohorte Maladie peu sévère (externe)	Entrée							Suivi externe						
Temps	Arrivée	J0	J2	J7	J14	J30	Congé	1M	3M	6M	12M	18M	24M	
Participant														
Sexe à la naissance	•													
Pays de naissance	•													
Statut vital	•	•	•	•	•	•	•	•	•	•	•	•	•	
Date du dernier statut vital connu	•	•	•	•	•	•	•	•	•	•	•	•	•	
Date du décès	•	•	•	•	•	•	•	•	•	•	•	•	•	
Heure du décès	•	•	•	•	•	•	•	•	•	•	•	•	•	
Lieu du décès	•	•	•	•	•	•	•	•	•	•	•	•	•	
Cause du décès	•	•	•	•	•	•	•	•	•	•	•	•	•	
Participant pédiatrique														
Enfant de moins d'un an	•													
Poids à la naissance	•													
Gestation	•													
Volet pédiatrique														
Patiente enceinte ?	•													
Date d'accouchement prévue	•													
Post partum (accouchement dans la dernière année)	•													
Résultat de la grossesse	•													
Date d'accouchement	•													
Statut COVID du bébé	•													
Bébé testé pour les pathogènes infectieux de la mère	•													
Résultat	•													
Méthode	•													
Type de participant														
Type de participant	•													
Diagnostic principal (ou primaire) relatif à l'hospitalisation	•													
Est-il un travailleur du milieu de la santé	•													
Est-il un travailleur dans un laboratoire de microbiologie	•													
Vit	•													
Vit avec	•													
Données démographiques														
Âge au moment de la visite	•	•	•	•	•	•	•	•	•	•	•	•	•	
Taille	•													
Poids	•													
IMC	•													
Tabagisme et consommation de drogues														
Statut tabagique	•													
Cigarette électronique	•													
Commentaire sur le tabagisme	•													
Usage de drogues	•													
Type de drogue	•													
Vaccination														
Vacciné														•
Nombre de doses														•
Type de vaccin														•
Dates														•
Effets secondaires														•
Signes vitaux et évaluation quotidienne														
Température	•	•	•	•	•	•								
Fréquence respiratoire	•	•	•	•	•	•								
Fréquence cardiaque	•	•	•	•	•	•								

Saturation d'O_2 à l'air ambiant	•	•	•	•	•	•							
Oxygénothérapie	•	•	•	•	•	•							
Saturation d'O_2 en oxygénothérapie	•	•	•	•	•	•							
FiO_2	•	•	•	•	•	•							
Pression systolique / diastolique	•	•	•	•	•	•							
Échelle AVPU		•	•	•	•	•							
Score de Glasgow (GCS/15)		•	•	•	•	•							
Flot urinaire en 24h		•	•	•	•	•							
Symptômes documentés													
Asymptomatique	•							•	•	•	•	•	•
Date d'apparition de(s) premiers(s) symptômes	•							•	•	•	•	•	•
Fièvre	•							•	•	•	•	•	•
Maux de gorge	•							•	•	•	•	•	•
Toux	•							•	•	•	•	•	•
Dyspnée	•							•	•	•	•	•	•
Hémoptysie/Sang dans l'expectoration	•							•	•	•	•	•	•
Rhinorrhée	•							•	•	•	•	•	•
Pharyngite	•							•	•	•	•	•	•
Mal d'oreilles	•							•	•	•	•	•	•
Arthralgie	•							•	•	•	•	•	•
Fatigue	•							•	•	•	•	•	•
Douleur thoracique	•							•	•	•	•	•	•
Douleur abdominale	•							•	•	•	•	•	•
Myalgie	•							•	•	•	•	•	•
Respiration sifflante	•							•	•	•	•	•	•
Nausée / vomissement	•							•	•	•	•	•	•
Maux de tête	•							•	•	•	•	•	•
Confusion	•							•	•	•	•	•	•
Diarrhée	•							•	•	•	•	•	•
Convulsions	•							•	•	•	•	•	•
Conjonctivite	•							•	•	•	•	•	•
Démangeaisons de la peau	•							•	•	•	•	•	•
Anorexie	•							•	•	•	•	•	•
Oedème des jambes	•							•	•	•	•	•	•
Trouble de la parole (Aphagie/dysphasie)	•							•	•	•	•	•	•
Perte du goût/perte de l'odorat	•							•	•	•	•	•	•
Faiblesse ou engourdissement aux extrémités	•							•	•	•	•	•	•
Antécédents médicaux													
AVC antérieur	•												
Accident ischémique transitoire (AIT) antérieur	•												
Fibrillation auriculaire ou flutter	•												
Autre problème neurologique (autre que AVC/AIT)	•												
Infarctus du myocarde antérieur	•												
Maladie coronarienne	•												
Insuffisance cardiaque	•												
Fraction d'éjection du ventricule gauche de base pré-hospitalisation	•												
Autre problème cardiaque chronique	•												
Maladies hématologiques chroniques	•												
Diabète	•												
Obésité	•												
Hypertension	•												
Asthme	•												
Problèmes rhumatologiques	•												
Problèmes rénaux chroniques	•												
Dialyse	•												

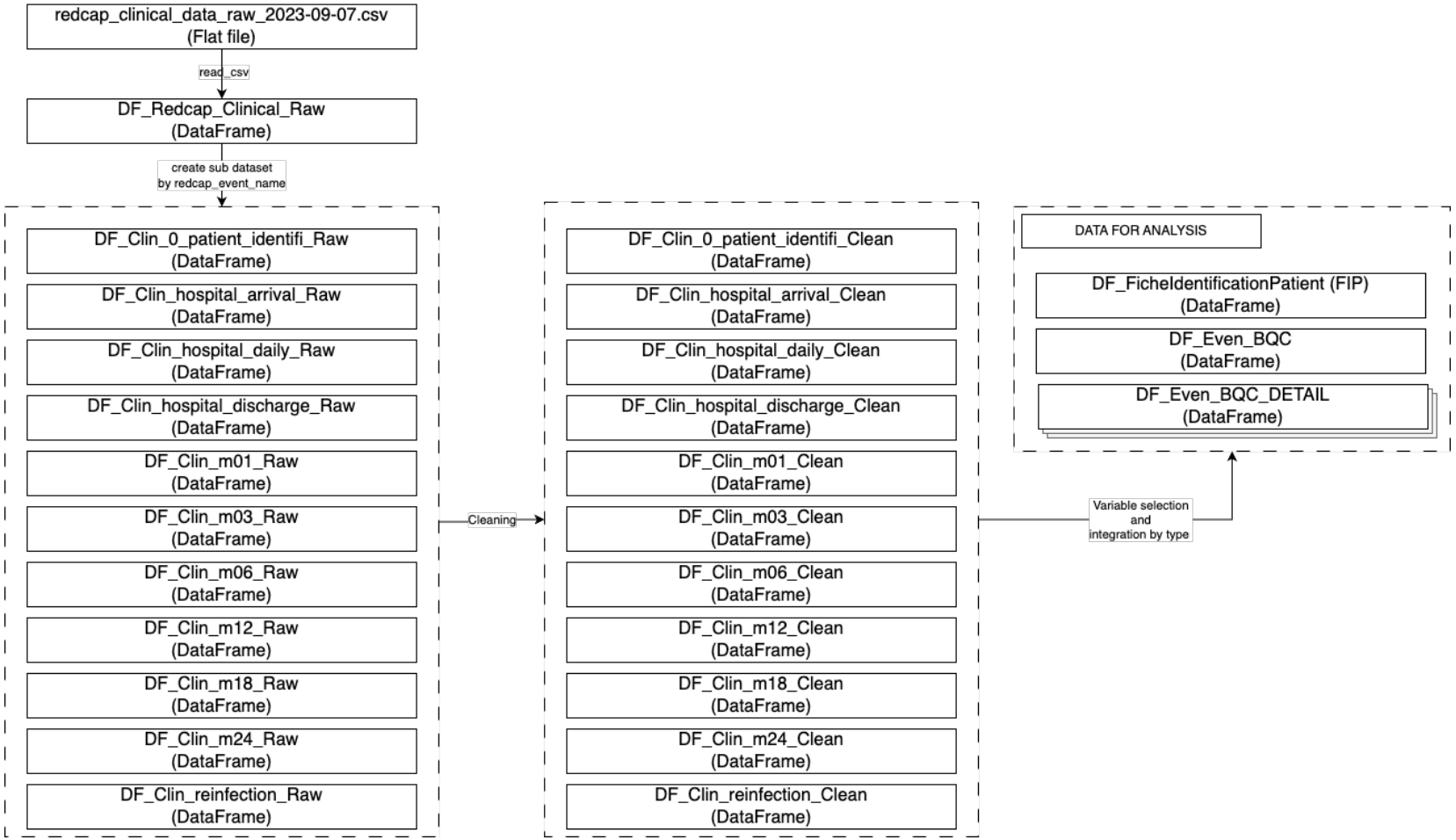
MPOC: Bronchite chronique, emphysème	•													
Sévérité de la MPOC	•													
Hypertension pulmonaire	•													
Autres problèmes pulmonaires chroniques	•													
Démence	•													
Problèmes psychiatriques	•													
Maladie hépatique chronique	•													
Malnutrition	•													
Cancer	•													
Type de cancer	•													
Avec métastase(s)	•													
Sous traitement actuellement	•													
HIV	•													
Immunosupprimé	•													
Autres antécédents med.	•													
Médication prise à la maison														
Inhibiteurs de l'enzyme de conversion	•													
Antagoniste des récepteurs de l'angiotensine														
Stéroïdes systémiques	•													
Anticoagulants	•													
Colchicine	•													
Autre médication immunosuppressive	•													
Analyses de laboratoire														
Décompte des globules blancs		•	•	•	•	•	max							
Décompte des Neutrophiles		•	•	•	•	•	max							
Décompte des Lymphocytes		•	•	•	•	•	min							
Décompte des Monocytes		•	•	•	•	•	max							
Décompte des Éosinophiles		•	•	•	•	•	max							
Décompte des Basophiles		•	•	•	•	•	max							
Décompte des plaquettes		•	•	•	•	•	min							
Mesure de l'hémoglobine		•	•	•	•	•	min							
Urée		•	•	•	•	•	max							
Créatinine		•	•	•	•	•	max							
NT-proBNP		•	•	•	•	•	max							
BNP		•	•	•	•	•	max							
Sodium Na+		•	•	•	•	•	max							
Potassium K+		•	•	•	•	•	max							
Protéine C-réactive (CRP)		•	•	•	•	•	max							
LDH		•	•	•	•	•	max							
CPK (CK)		•	•	•	•	•	max							
Albumine		•	•	•	•	•	min							
AST		•	•	•	•	•	max							
ALT		•	•	•	•	•	max							
Procalcitonine (PCT)		•	•	•	•	•	max							
Troponine T hs		•	•	•	•	•	max							
Troponine I hs		•	•	•	•	•	max							
Troponine T		•	•	•	•	•	max							
Troponine I		•	•	•	•	•	max							
APTT		•	•	•	•	•	max							
International Normalized Ratio (INR)		•	•	•	•	•	max							
Triglycérides		•	•	•	•	•	max							
Bilirubine totale		•	•	•	•	•	max							
Bilirubine directe (conjuguée)		•	•	•	•	•	max							
Glucose		•	•	•	•	•	max							
Lactate Veineux		•	•	•	•	•	max							
D-Dimère		•	•	•	•	•	max							

Fibrinogène		•	•	•	•	•	max						
Ferritine		•	•	•	•	•	max						
IL-6		•	•	•	•	•	max						
CD4		•	•	•	•	•	min						
CD8		•	•	•	•	•	min						
Support													
Support respiratoire		•	•	•	•	•	•						
Oxygénothérapie par canule / masque		•	•	•	•	•	•						
Support non invasif par LNHD		•	•	•	•	•	•						
Support non invasif CPAP/BPAP		•	•	•	•	•	•						
Support invasif avec ventilation mécanique		•	•	•	•	•	•						
Nombre de jours sous ventilation mécanique							•						
SpO_2 (plus basse associée avec le support le plus élevé)		•	•	•	•	•	•						
FiO_2 (en lien avec SpO_2)		•	•	•	•	•	•						
Thérapie de support		•	•	•	•	•	•						
Vasopresseurs / inotropes		•	•	•	•	•	•						
Nombre de jours sous							•						
Positionnement ventral		•	•	•	•	•	•						
iNO		•	•	•	•	•	•						
ECMO		•	•	•	•	•	•						
Support invasif HFOV (High-frequency oscillatory ventilation)		•	•	•	•	•	•						
Trachéostomie		•	•	•	•	•	•						
Transfusion de produits sanguins		•	•	•	•	•	•						
Agents neuro-bloquants		•	•	•	•	•	•						
Dialyse / hémofiltration		•	•	•	•	•	•						
Autre, préciser		•	•	•	•	•	•						
Score de fragilité													
Score de fragilité		•					•	•	•	•	•	•	•	•
Résumé hospitalier													
Date et heure de l'arrivée à l'hôpital							•						
Visite à l'urgence seulement							•						
Est-ce un transfert d'un autre établissement ?							•						
Nom de l'établissement							•						
Si transféré d'un autre hôpital, date d'admission initiale							•						
Date d'admission aux soins intensifs							•						
Date de congé des soins intensifs							•						
Date de congé de l'hôpital							•						
Disposition							•						
Statut lors du congé							•						
Si COVID positif, quel est le degré de sévérité (selon WHO) le plus sévère atteint?							•						
Habilité à prendre soin de soi au moment du départ de l'hôpital vs pré-COVID							•						
Dernier niveau de soin							•						
Diagnostic COVID et test de pathogènes													
Date du test diagnostic pour COVID19							•						
Résultat							•						
Type d'échantillon							•						
Autres infections virales documentées durant l'hospitalisation							•						
Adenovirus							•						
Influenza							•						
RSV							•						

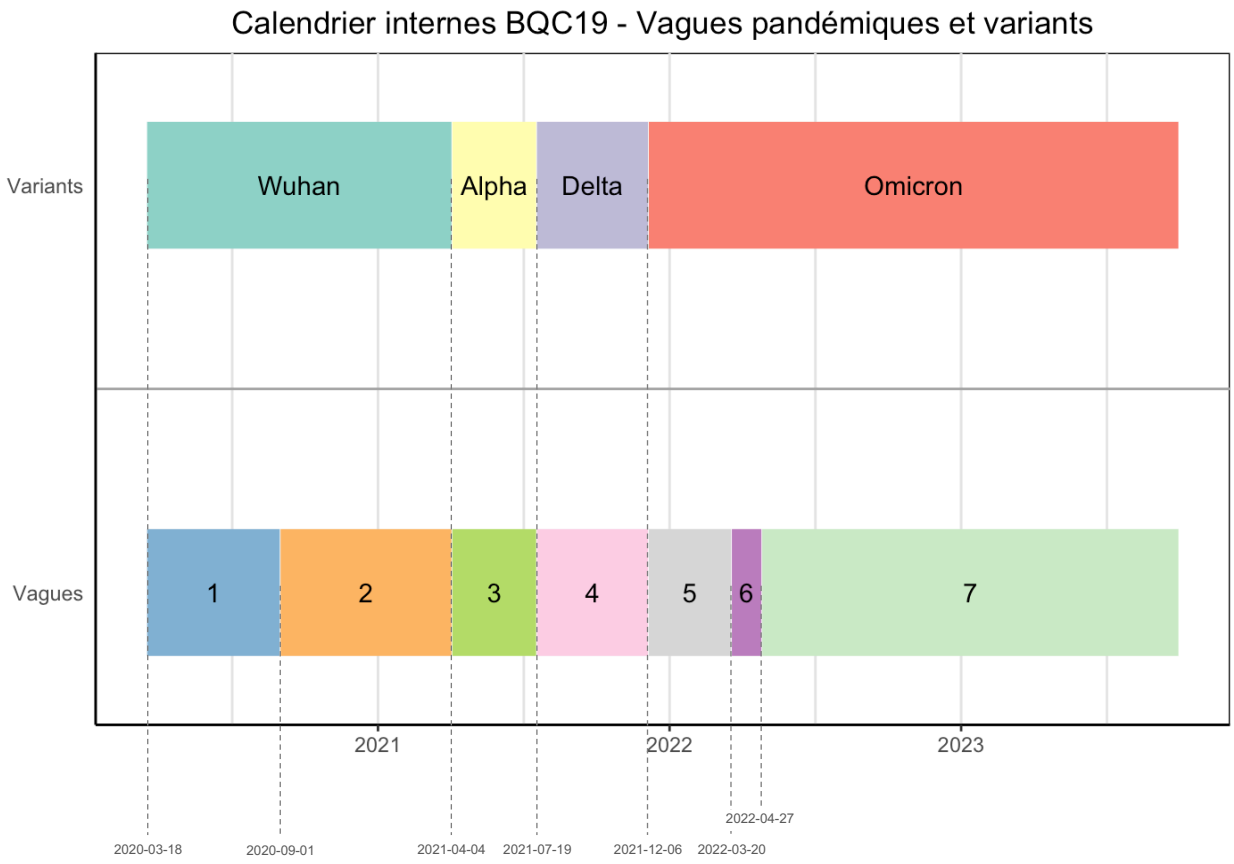
Parainfluenza							•						
Rhinovirus / Enterovirus							•						
Metapneumovirus							•						
Autre, préciser							•						
Autres infections bactériennes documentées durant l'hospitalisation							•						
Complications durant l'hospitalisation													
SDRA							•						
Spécifier la sévérité							•						
Pneumonie bactérienne							•						
Pneumonie virale							•						
Bronchiolite							•						
Pneumonie organisée cryptogénique (POC)							•						
Embolie pulmonaire							•						
Thrombose veineuse profonde (DVT)							•						
Pneumothorax							•						
Épanchement pleural							•						
Fonction ventriculaire gauche							•						
LVEF							•						
Insuffisance cardiaque décompensée							•						
Myocardite							•						
Endocardite							•						
Péricardite							•						
Fibrillation auriculaire(FA)/Flutter							•						
Fibrillation/Tachycardie ventriculaire							•						
Autre type d'arythmie cardiaque							•						
Spécifier							•						
Arrêt cardiaque							•						
Ischémie cardiaque STEMI							•						
Ischémie cardiaque NSTEMI							•						
Accident vasculaire cérébral							•						
Pancréatite							•						
Accident ischémique transitoire							•						
Convulsions							•						
Méningite / Encéphalite							•						
Rhabdomyolyse / myosite							•						
Trouble de la coagulation							•						
coagulation intravasculaire disséminée													
Bactériémie							•						
Anémie							•						
Hémorragie gastro-intestinale							•						
Dysfonction hépatique							•						
Insuffisance rénale aigue							•						
Hyperglycémie							•						
Hypoglycémie							•						
Autre							•						
Spécifier							•						
Autres tests effectués durant l'hopitalisation													
Radiographie des poumons							•						
Tomodensitométrie thorax							•						
Tomodensitométrie de l'abdomen							•						
Tomodensitométrie de la tête							•						
ECG							•						
Échographie au point d'intervention (POCUS)							•						
Échocardiogramme							•						
Échographie cardiaque													

Angiographie coronarienne (Cathétérisme cardiaque)						•							
Intervention coronarienne percutané(stent)						•							
Autres tests d'imagerie						•							
Médication durant l'hospitalisation													
Remdesivir						•							
Ribavirin						•							
Lopinavir/Ritonavir ("Kaletra")						•							
Interferon alpha						•							
Interferon beta						•							
Neuraminidase inh						•							
Autre antiviral						•							
Préciser						•							
Azithromycin ("Zithromax")						•							
Autre antiobiotique						•							
Antifongique						•							
Corticostéroide systémique						•							
Hydroxychloroquine ("Plaquenil")						•							
Chloroquine ("Aralen")						•							
Ivermectin ("Stromectol")						•							
Tocilizumab ("Actemra")						•							
Sarilumab ("Kevzara")						•							
Kineret ("Anakinra")						•							
Autre immunomodulateur						•							
Colchicine						•							
Plasma convalescent						•							
Cellules souches / cellules T éduquées						•							
IVIG						•							
Autre traitement pour COVID-19						•							
Suivi													
Visite à l'hôpital suite au congé							•	•	•	•	•	•	•
Date admission							•	•	•	•	•	•	•
Établissement							•	•	•	•	•	•	•
Cause							•	•	•	•	•	•	•
Statut fonctionnel													
Mobilité							•	•	•	•	•	•	•
Soins personnels							•	•	•	•	•	•	•
Activités habituelles							•	•	•	•	•	•	•
Douleur et inconfort							•	•	•	•	•	•	•
Anxiété et dépression							•	•	•	•	•	•	•
Essoufflement							•	•	•	•	•	•	•
État de santé							•	•	•	•	•	•	•
Quelle difficulte éprouvez-vous ...													
à soulever ou porter une charge de 10 lbs?							•	•	•	•	•	•	•
à marcher à travers d'une piece?							•	•	•	•	•	•	•
à transférer d'une chaise a un lit?							•	•	•	•	•	•	•
à monter 1 étages à pied?							•	•	•	•	•	•	•
Combien de fois êtes-vous tombé dans la dernière année?							•	•	•	•	•	•	•

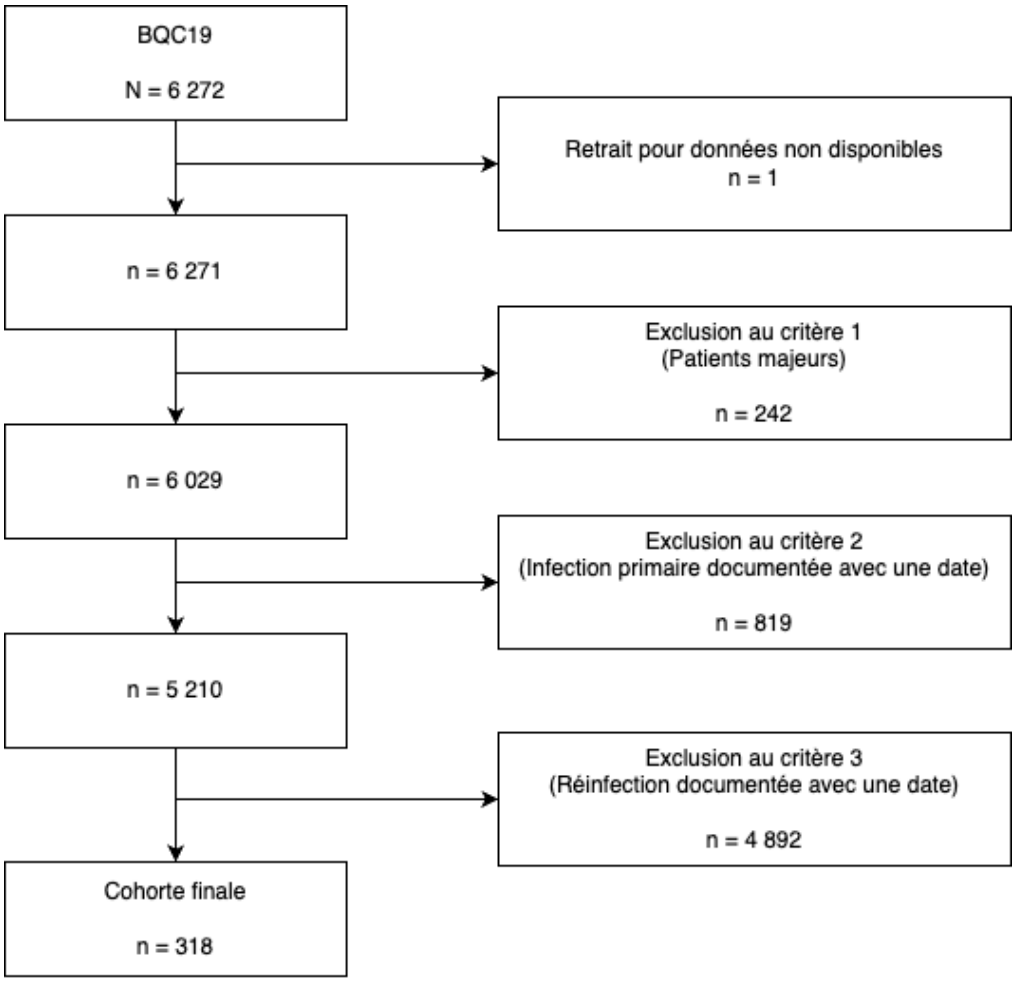
Annexe 3 Schéma de la transformation globale des données



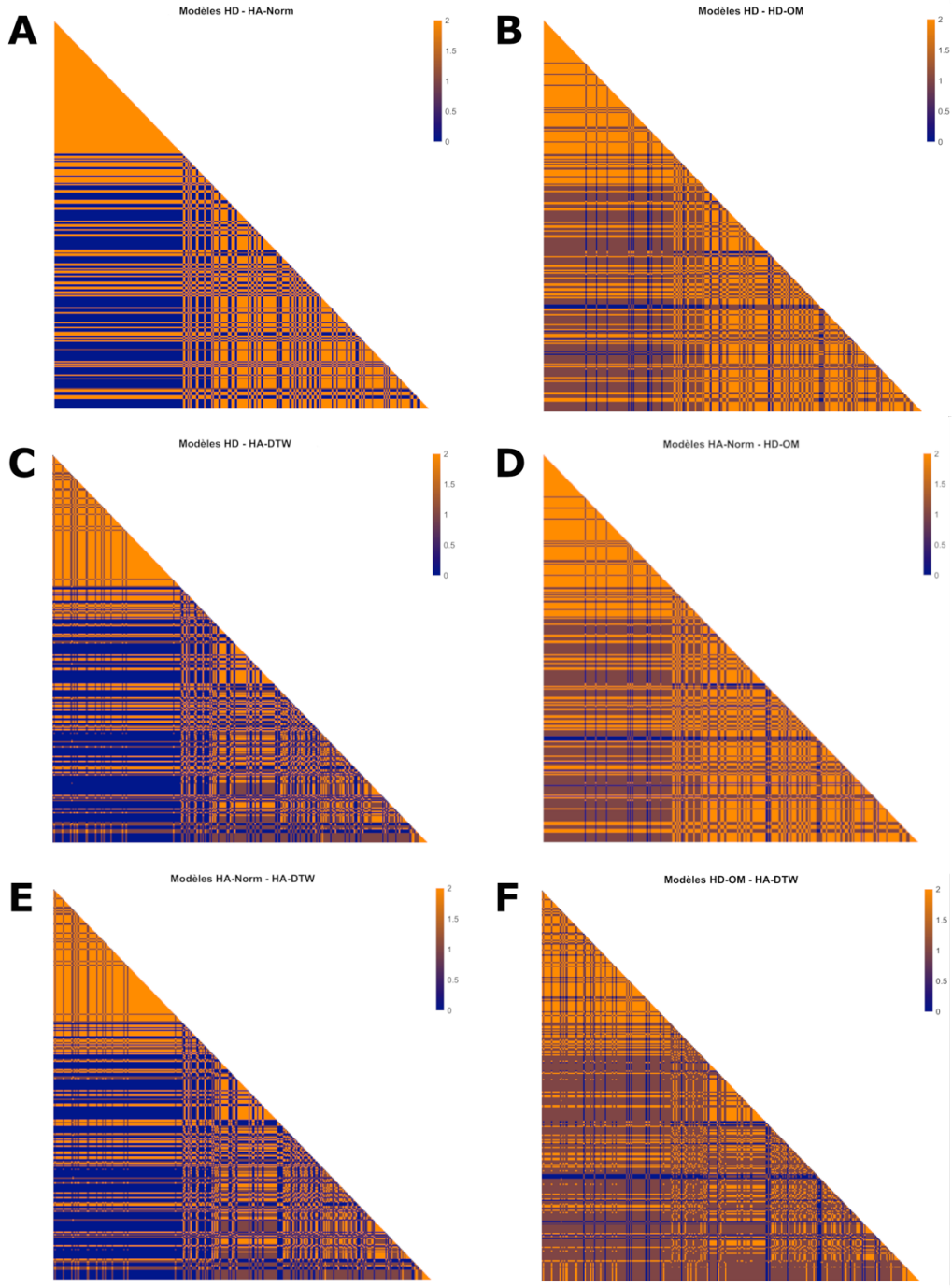
Annexe 4 Calendriers internes à la BQC19



Annexe 5 Diagramme de flux d'inclusion des participants à la cohorte



Annexe 6 Cartes de chaleur du nombre d'occurrence où chaque paire d'individus est groupé ensemble, modèle par modèle



Annexe 7 Reconnaissance de l'approbation éthique

HEC MONTRÉAL

Comité d'éthique de la recherche

Le 11 juillet 2023

À l'attention de :
Jean-Frédéric Boulianne
HEC Montréal

Objet : Reconnaissance de l'approbation éthique de votre projet de recherche

 # Projet : 2024-5571

Titre du projet de recherche : Impacts de l'immunité hybride au SRAS-CoV-2 dans la Biobanque Québécoise de la COVID-19 (BQC19) : Apport de l'apprentissage machine pour l'étude des facteurs, des déterminants de la santé et des trajectoires dans le réseau sociosanitaire (titre provisoire)

Source de financement : Agence de santé publique du Canada

Bonjour Jean-Frédéric,

Votre projet de recherche a fait l'objet d'une reconnaissance de l'approbation éthique de votre projet de recherche émise par le CER du Centre universitaire de santé McGill en vertu de *l'Entente pour la reconnaissance des certificats d'éthique des projet de recherche à risque minimal* entre les universités québécoises.

Prenez note que cette reconnaissance est **valide jusqu'au 1 août 2024.**

Vous devrez obtenir le renouvellement de votre reconnaissance à l'aide du formulaire *F7 - Renouvellement annuel*. Un rappel automatique vous sera envoyé par courriel quelques semaines avant l'échéance de la certification. Si votre projet est terminé vous devrez remplir le formulaire *F9 - Fin de projet.*

Notez qu'en vertu de la *Politique relative à l'éthique de la recherche avec des êtres humains de HEC Montréal*, il est de la responsabilité des chercheurs d'assurer que leurs projets de recherche conservent une approbation éthique pour toute la durée des travaux de recherche et d'informer le CER de la fin de ceux-ci. De plus, toutes modifications significatives du projet doivent être transmises au CER avant leurs applications.

Vous pouvez procéder à la collecte de données pour laquelle ce certificat a été délivré.

Nous vous souhaitons bon succès dans la réalisation de votre recherche.

Le CER de HEC Montréal

HEC MONTRÉAL

Comité d'éthique de la recherche

Le 19 août 2024,

À l'attention de : Jean-Frédéric Boulianne, HEC Montréal

Objet : Reconnaissance du renouvellement de l'approbation éthique par un autre CER

 # Projet : 2024-5571

Titre du projet de recherche : Impacts de l'immunité hybride au SRAS-CoV-2 dans la Biobanque Québécoise de la COVID-19 (BQC19) : Apport de l'apprentissage machine pour l'étude des facteurs, des déterminants de la santé et des trajectoires dans le réseau sociosanitaire (titre provisoire)

Bonjour,

Pour donner suite à la réception du renouvellement du certificat d'approbation éthique émis par le CER de l'Université McGill, le CER de HEC Montréal reconnait celui-ci en vertu de *l'Entente pour la reconnaissance des certificats d'éthique des projets de recherche à risque minimal.*

Ce renouvellement est valide jusqu'au 01 août 2025.

Nous vous souhaitons bon succès dans la poursuite de votre recherche.

Cordialement,

Le CER de HEC Montréal

Maurice Lemelin
Président
CER de HEC Montréal

Signé le 2024-08-19 à 13:53

HEC MONTRÉAL

Comité d'éthique de la recherche

ATTESTATION DE RECONNAISSANCE D'APPROBATION ÉTHIQUE COMPLÉTÉE

La présente atteste que le projet de recherche décrit ci-dessous a fait l'objet des approbations en matière d'éthique de la recherche avec des êtres humains nécessaires selon les exigences de HEC Montréal.

La période de validité de la reconnaissance du certificat d'approbation éthique émis pour ce projet est maintenant terminée. Si vous devez reprendre contact avec les participants ou reprendre une collecte de données pour ce projet, la certification éthique doit être réactivée préalablement. Vous devez alors prendre contact avec le secrétariat du CER de HEC Montréal.

Projet # : 2024-5571 - Mémoire Jean-Frédéric Boulianne (essai 2)

Titre du projet de recherche : Impacts de l'immunité hybride au SRAS-CoV-2 dans la Biobanque Québécoise de la COVID-19 (BQC19) : Apport de l'apprentissage machine pour l'étude des facteurs, des déterminants de la santé et des trajectoires dans le réseau sociosanitaire (titre provisoire)

Chercheur principal : Jean-Frédéric Boulianne

Cochercheurs : Delphine Bosson-Rieutort; Simon Rousseau

Directeur/codirecteurs : Denis Larocque

Date d'approbation initiale du projet : 11 juillet 2023

Date de fermeture de l'approbation éthique : 12 novembre 2024

Maurice Lemelin
Président
CER de HEC Montréal