COVID-19 : Les facteurs qui influencent sa sévérité

Statistics 2

Amani Mahdi (CMB) & Jenna Ouatah (BIP)

1 Introduction

Selon l'Organisation Mondiale de la Santé (OMS), la COVID-19 est une maladie infectieuse causée par le virus SARS-CoV-2. La plupart des personnes infectées présentent une forme respiratoire légère à modérée. Ainsi, cette étude vise à mieux comprendre les facteurs intrinsèques pouvant influencer la gravité de la COVID-19.

Nous avons donc posé la question centrale suivante : **"Quels facteurs influencent la sévérité de la COVID-19 ?"**

Pour cela, notre analyse s'est articulée autour de deux axes principaux :

Dans un premier temps, nous avons choisi d'évaluer la sévérité du COVID-19 travers le rapport SpO2/FiO2, un indicateur clé de la fonction respiratoire. Ce ratio reflète l'efficacité des échanges gazeux au niveau pulmonaire. Une valeur normale est d'environ 452, tandis qu'un rapport inférieur à 300 indique une altération sévère des échanges respiratoires. Ainsi, un ratio faible est associé à une détresse respiratoire plus marquée. Dans cette partie nous avons analysé le lien entre ce critère de sévérité et divers facteurs cliniques, tel que les paramètres biologiques, les comorbidités associés, et la prise d'antiviraux.

Dans un second temps, nous avons cherché à évaluer la sévérité du COVID-19 à travers l'expression génique de patients appartenant à trois groupes : sains, atteints de COVID-19 léger ou de COVID-19 sévère. Notre but étant d'étudier les différences d'expression génique entre ces profils, afin d'identifier des gènes dont l'expression varie selon la gravité de l'infection.

2 Étude clinique du COVID-19

2.1 Analyse exploratoire des données

L'objectif de cette partie est d'évaluer si le ratio SpO2/FiO2, utilisé comme critère de sévérité de la COVID-19, est influencé par différents facteurs cliniques. Plus précisément, nous avons étudié les liens entre ce ratio et : (1) la présence de comorbidités, (2) les paramètres biologiques et (3) la prise d'agents antiviraux.

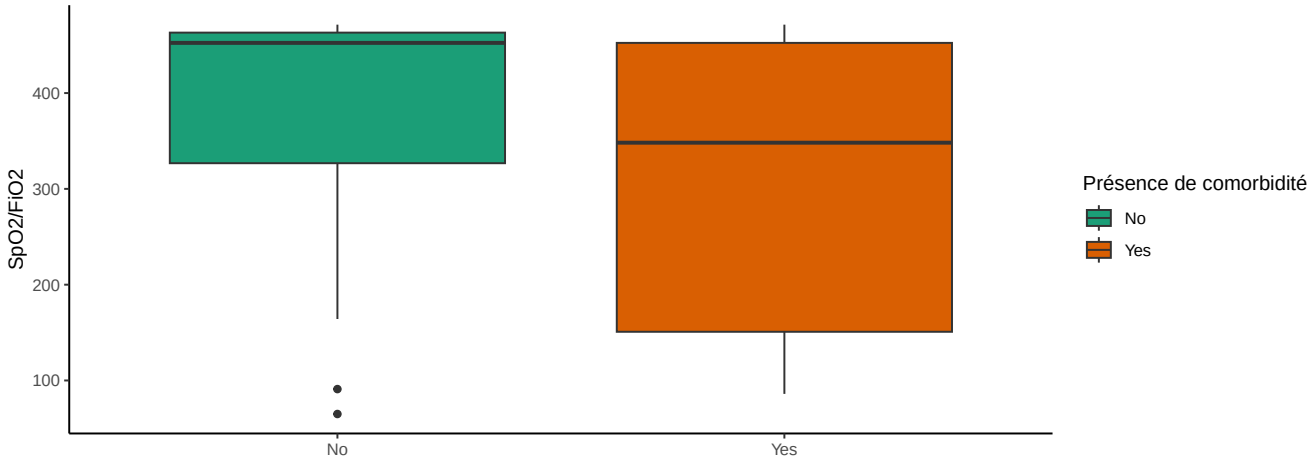
2.1.1 Étude du ratio SpO2/FiO2 en fonction des comorbidités

L'analyse du ratio SpO2/FiO2 en fonction des comorbidités vise à explorer leur impact potentiel sur la sévérité de la COVID-19. Parmi ces comorbidités, la BPCO (Bronchopneumopathie Chronique Obstructive) est une pathologie respiratoire connue pour aggraver l'hypoxémie. L'hypertension artérielle et les maladies cardiaques altèrent la fonction vasculaire, pouvant ainsi impacter les échanges respiratoires. Le diabète est associé à des déséquilibres métaboliques susceptibles d'aggraver la réponse inflammatoire. Enfin, la maladie rénale chronique.

Pour évaluer l'impact des comorbidités sur le ratio SpO2/FiO2, un test de Wilcoxon a été réalisé afin de comparer les moyennes du ratio SpO2/FiO2 entre les patients présentant ou non certaines comorbidités. Des Boxplot ont également été créés pour visualiser les distributions. Les hypothèses étaient les suivantes : l'hypothèse nulle H0, absence de différence entre les moyennes et l'hypothèse alternative H1, existence d'une différence significative. Le critère de décision est basé sur la p-value : si $p < 0{,}05$, H0 est rejetée et H1 est conservée.

Un test de Shapiro-Wilk a été effectué afin de vérifier la normalité des données. Comme aucune variable ne suivait une loi normale, l'alternative non paramétrique au test t a été prévilégiée.

Figure 1. SpO2/FiO2 en fonction de la présence ou non de comorbidités

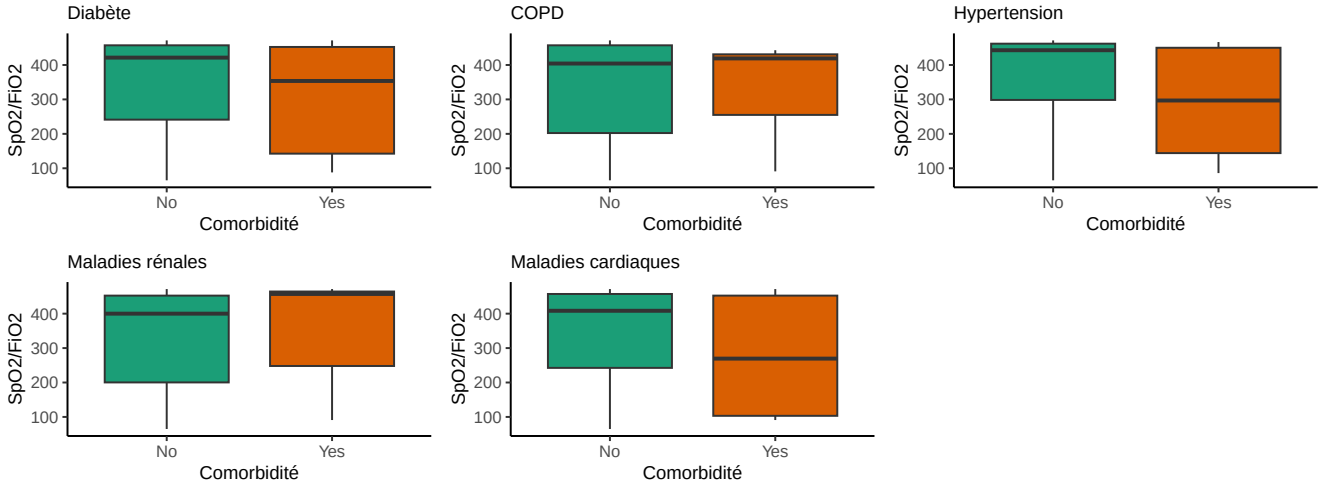


D'après la Figure 1, les patients présentant au moins une comorbidité semblent avoir un ratio SpO2/FiO2 plus faible que les autres. Cette différence est statistiquement significative (Tableau 1).

Tableau 1 : Résultats des tests de Wilcoxon

Parametère	p-value
Comorbidité	0.010382675
Diabète	0.286951767
COPD	0.493977128
Hypertension	0.008659563
Maladies cardiaques	0.239145473
Maladies rénales	0.123150193

Figure 2. Ratio SpO2/FiO2 en fonction des comorbidités



La Figure 2 montre que les patients hypertendus présentent un ratio SpO2/FiO2 plus faible que les non hypertendus avec un ratio médian d'environ $300mmHg$ chez les patients hypertendus contre $450mmHg$ chez les non hypertendus. Cette différence observée est statistiquement significative (p-value $< 0,05$, Tableau 1) Cependant, pour le diabète, les maladies cardiaques, les maladies rénales chroniques ainsi que la BPCO, même si, certaines variations peuvent-être observées, ces dernières ne sont pas statistiquement significatives (p-value $> 0,05$, Tableau 1).

2.1.2 Étude du ratio SpO2/FiO2 en fonction des données biologiques

L'analyse du ratio SpO2/FiO2 en lien avec divers paramètres biologiques permet d'évaluer la relation potentielle entre ces variables et la sévérité de la COVID-19. Parmi ces marqueurs, les lymphocytes et l'IL-6 sont des marqueurs de la réponse immunitaire, la CRP reflète l'inflammation, le D-dimère est un indicateur de l'activation de la coagulation, et des enzymes hépatiques telles que l'AST et la LDH témoignent de lésions cellulaires. Enfin, le DFG influence la fonction rénale.

Pour évaluer l'existence d'une corrélation entre le ratio SpO2/FiO2 et les paramètres biologiques, un test de corrélation a été réalisé. Des nuages de points ont également été créés pour visualiser données.

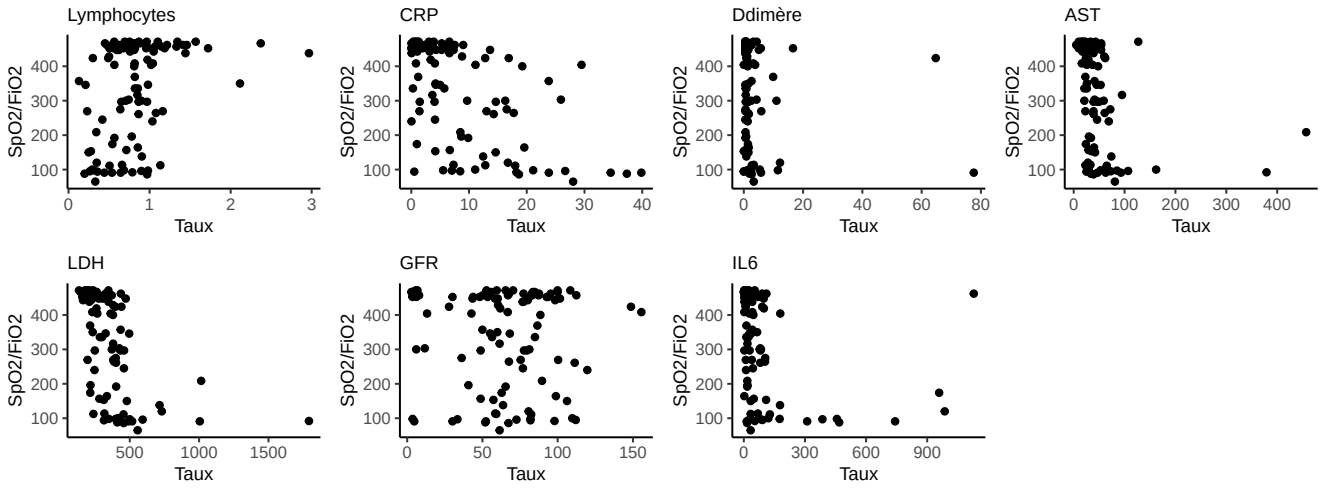
Tableau 2 : Résultats des tests de corrélation

Parametère	p-value
Lymphocytes	3.822767e-04
CRP	6.536771e-11
D-dimer	1.147363e-03
AST	1.133140e-07
LDH	3.539304e-13
GFR	4.890116e-01
IL6	5.419554e-06

Les hypothèses étaient les suivantes : H0, absence de corrélation entre les variables, et H1, existence d'une corrélation entre les variables.

La normalité des données a été vérifiée par un test de Shapiro-Wilk, comme aucune variable ne suivait une distribution normale, c'est le coefficient de corrélation de Kendall qui a été utilisé.

Figure 3. Ratio SpO2/FiO2 en fonction valeurs du bilan sanguin



Comme le montre les graphiques de la Figure 3, les parmètres du bilan sanguin ne semblent pas avoir de lien visible avec le ratio Sp02/Fi02. En effet, les points sont dispersés un peu partout.

Cependant, les tests de corrélation ont montré que les paramètres biologiques Lymphocytes, CRP, D-dimère, AST, LDH et IL-6, présentent une corrélation significative avec le ratio SpO2/FiO2. Seul, le taux de GFR, ne présente pas de lien statistiquement significatif (Tableau 2).

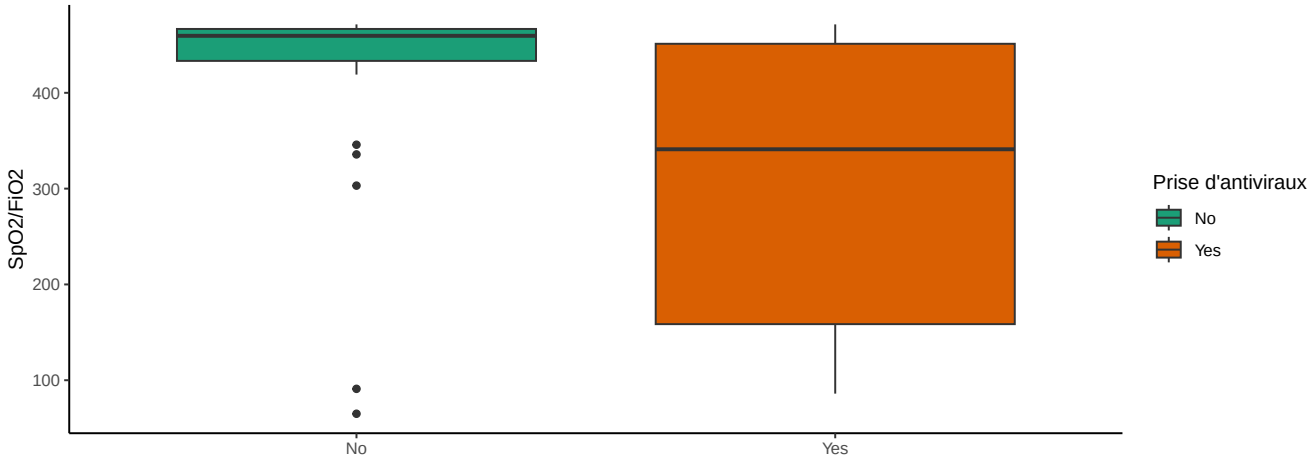
2.1.3 Étude du ratio SpO2/FiO2 en fonction des traitements

L'analyse du ratio SpO2/FiO2 en fonction des traitements vise à évaluer l'impact potentiel de certaines thérapies sur la sévérité de la COVID-19. Parmi ces traitements, les stéroïdes, administrés par inhalation

ou injection (stéroïdes systémiqyes), agissent de manière générale sur l'ensemble du système immunitaire et le Tocilizumab, qui agit de manière spécifique sur le système immunitaire en ciblant l'interleukine-6.

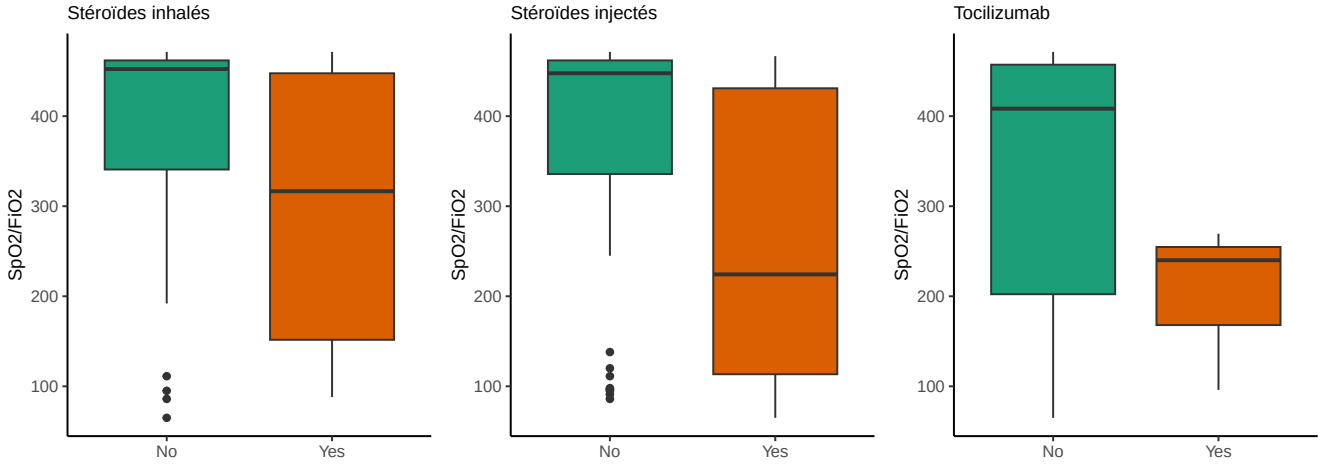
Pour évaluer l'impact des traitements sur le ratio SpO2/FiO2, un test de Wilcoxon a été réalisé afin de comparer les moyennes du ratio SpO2/FiO2 entre les patients prenants ou non certains antiviraux. Des Boxplot ont également été créés pour visualiser les distributions.

Figure 4. SpO2/FiO2 en fonction de la prise ou non d'antiviraux



D'après la Figure 4, les patients ayant reçu un traitement antiviral semblent avoir un ratio SpO2/FiO2 plus faible que les autres. Cette différence est statistiquement significative (Tableau 3).

Figure 5. Ratio SpO2/FiO2 en fonction des antivraux administrés



D'après la figure 5, les patients qui ont reçu un traitement antiviral semblent avoir un ratio SpO2/FiO2 plus bas. Cependant, cette différence n'est statistiquement significative uniquement pour la prise de stéroïdes, qu'ils soient administrés par inhalation ou par injection. Les patients ayant reçu du Tocilizumab, quant à eux, ne possèdent pas un ratio SpO2/FiO2 significativement différent de ceux qui n'en ont pas reçu (Tableau 3).

Tableau 3 : Résultats des tests de Wilcoxon

Parametère	p-value
Antiviraux	2.558860e-04
Stéroides inhalés	2.061764e-03
Stéroides injectés	1.705224e-05
Tocilizumab	9.782905e-02

En conculsion, l'étude clinique a permis de mettre en évidence plusieurs facteurs associés à une baisse significative du ratio SpO2/FiO2, indicateur clé de la sévérité de la COVID-19. Les paramètres biologiques impliqués dans la réponse immunitaire (lymphocyte, IL-6), dans l'inflammation (CRP), et dans les lésions cellulaires (AST, LDH) sont significativement corrélés à une diminution du rapport SpO2/FiO2. De plus, l'hypertension, connu pour de favoriser les complications cardiaques et respiratoires, est également associée à une baisse significative du rapport SpO2/FiO2. Concernant les stéroïdes, bien qu'ils soient également associés à une baisse du rapport SpO2/FiO2, cette observation pourrait s'expliquer par le fait qu'ils ont probablement été administrés lorsque les patients présentaient déjà une atteinte respiratoire sévère. Leur effet étant retardé, ils visent principalement à stabiliser l'inflammation afin d'éviter son aggravation, ce qui justifie l'absence d'amélioration immédiate du rapport.

2.2 Analyse multivariée

Après avoir analysé la relation entre le ratio SpO2/FiO2 et la présence de comorbidités, les données du bilan sanguin ainsi que les traitements antiviraux, nous avons réalisé une régression pour prédire ce ratio.

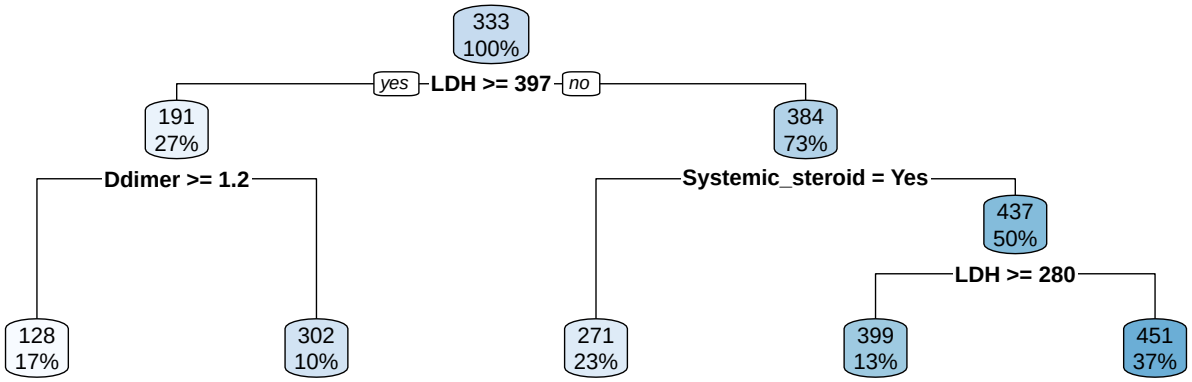
Pour ce faire, nous avons crée deux échantillons : le premier, composé de 80 % des données, a servi à l'entraînement du modèle, tandis que le second, représentant les 20 % restants, a été utilisé pour tester objectivement le modèle.

Dans un premier temps, nous avons défini notre modèle en prenant le ratio SpO2/FiO2 comme variable de réponse. Dans un second temps, nous avons utilisé ce modèle pour prédire ce ratio sur l'échantillon de test.

Dans cette analyse, nous avons d'abord construit un arbre de régression (1), puis une forêt de régression (2).

2.2.1 Arbre de régression

Figure 6. Arbre de régression − Ratio SpO2/FiO2

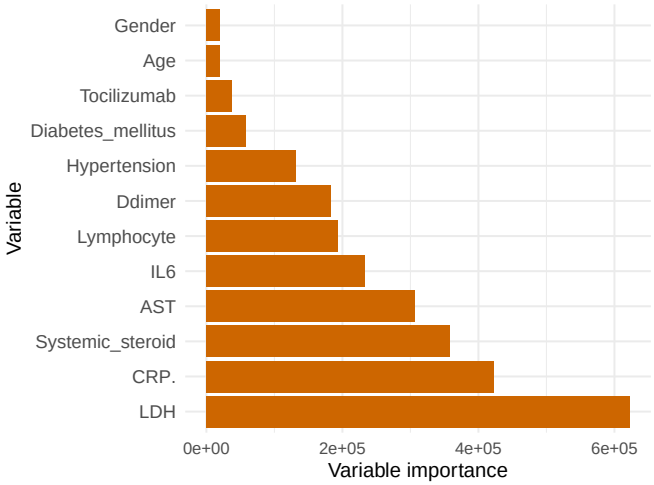


Comme l'illustre la figure 6, l'arbre de décision se divise en trois branches basées sur le taux de LDH, les D-dimères et l'administration de stéroïdes injectés.

L'analyse de l'importance des variables (figure 7) révèle que le taux de LDH, l'AST, CRP ainsi que la prise de stéroïdes systémiques ont le plus d'influence sur le modèle, avec des valeurs d'importance respectives de 6.10^5, 4.10^5, $3,5.10^5$ et $2,8.10^5$.

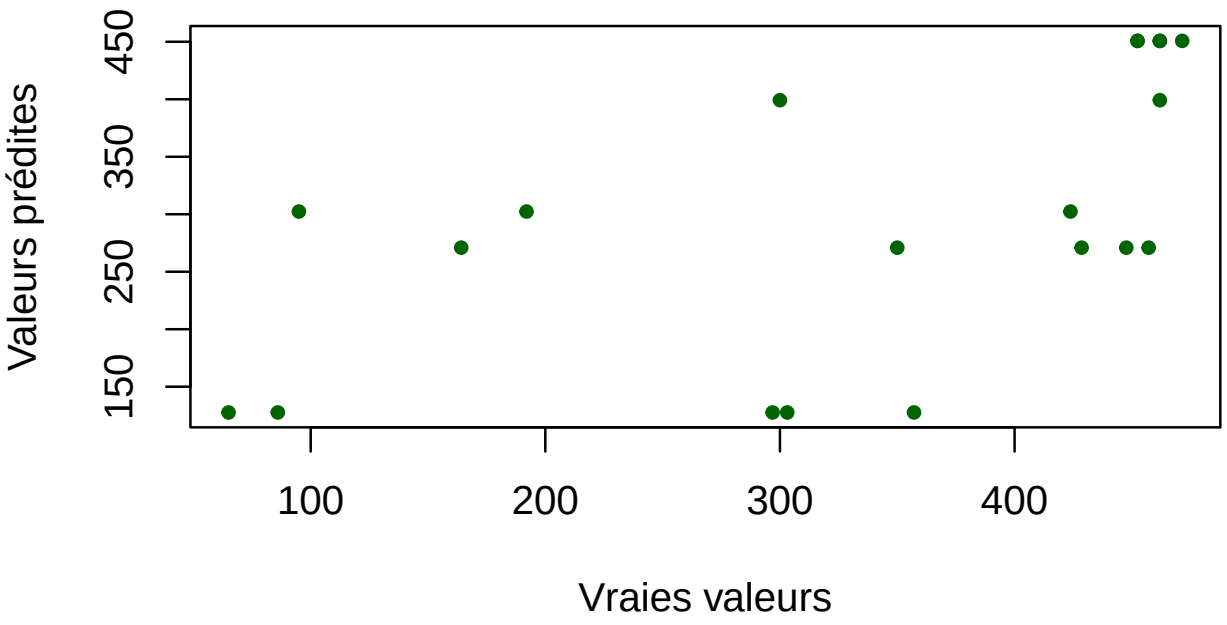
Toutefois, le modèle a une performance limitée dans la prédiction du ratio, puisque les points sont très dispersés sur le graphique représentant les valeurs prédites en fonction des vraies valeurs de l'échantillon (figure 8). Ainsi, il n'est pas possible, à ce stade, de formuler une hypothèse claire sur les paramètres ayant le plus d'impact sur le ratio SpO2/FiO2.

Figure 7. Importance des variables



	Variable	Importance
1	LDH	622456.85
2	CRP.	423097.35
3	Systemic_steroid	357733.55
4	AST	306359.48
5	IL6	232560.72
6	Lymphocyte	193082.76
7	Ddimer	182802.49
8	Hypertension	131796.57
9	Diabetes_mellitus	58306.05
10	Tocilizumab	37656.16
11	Age	19435.35
12	Gender	19435.35

Figure 8. Valeurs prédites en fonction des vraies valeurs



2.2.2 Forêt de régression

Pour améliorer les résultats précédents, nous avons construit une forêt de régression, puis avons comme précédemment testé notre modèle sur l'échantillon test.

La figure 10 montre que les points suivent généralement une tendance linéaire, bien que celle-ci soit plus marquée aux extrémités. Cela indique une meilleure performance du modèle, qui parvient à prédire plus précisément les valeurs du ratio SpO2/FiO2.

L'analyse des variables importantes (figure 9) met en lumière les mêmes facteurs que précédemment : le taux de LDH, de CRP, d'AST et l'administration de stéroïdes systémiques. Ces résultats suggèrent que ces variables ont le plus d'effet sur le ratio SpO2/FiO2. Il conviendait néanmoins de garder à l'esprit que ce type de régression ne permet pas de tirer des conclusions sur la significativité statistique de ces résultats.

Figure 9. Importance des variables

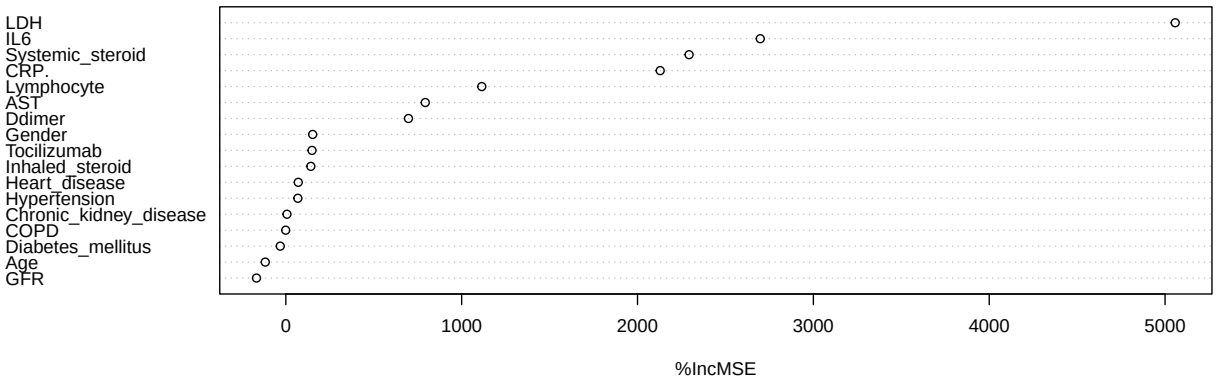
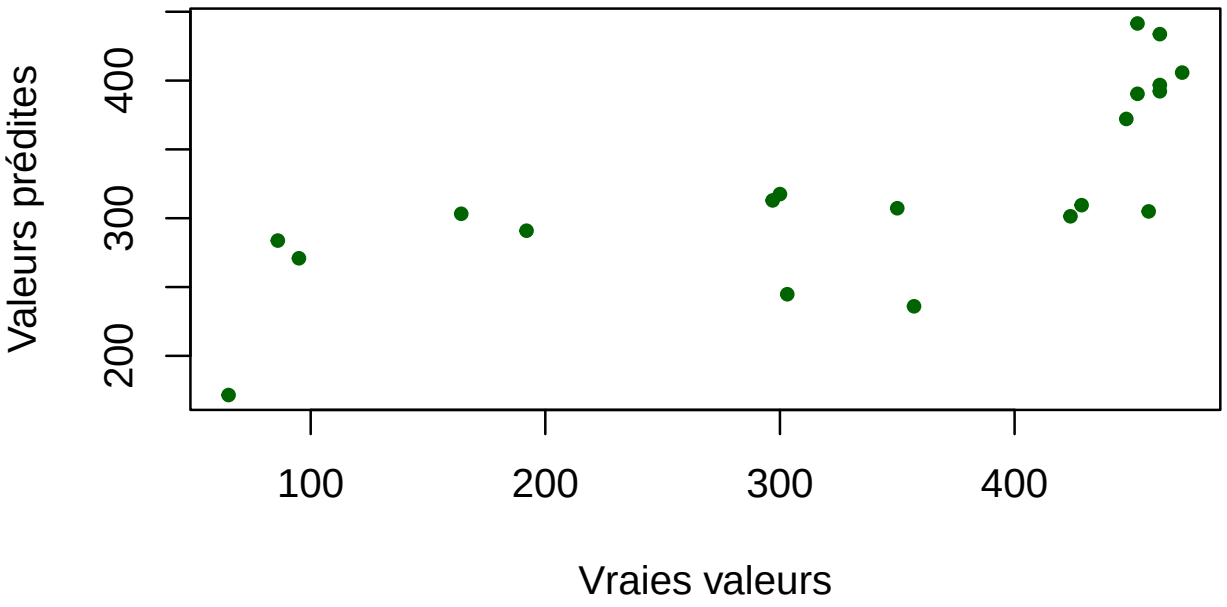


Figure 10. Valeurs prédites en fonction des vraies valeurs



3 Étude transcriptomique du COVID-19

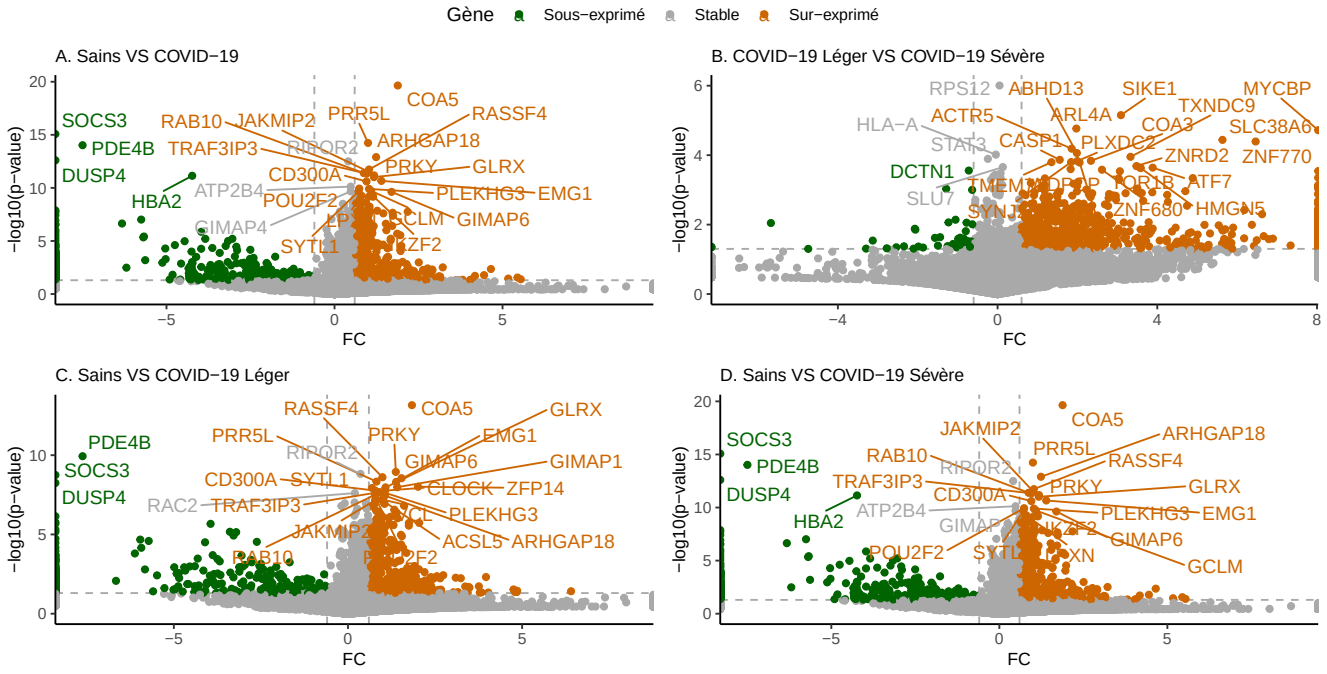
Nous étudions ici les données issues de RNA-Seq dans les cellules NK, afin d'examiner s'il existe des gènes qui s'expriment différemment en fonction de la sévérité du COVID-19 (Sains/COVID-19 Léger/COVID-19 Sévère).

3.1 Analyse exploratoire des données

Afin de visualiser l'expression des gènes, nous avons créé quatre Volcano-plots, correspondant à quatre situations distinctes, où chaque fois nous avons comparé deux groupes de patients : Sains vs COVID-19, COVID-19 Léger vs COVID-19 Sévère, Sains vs COVID-19 Léger, et enfin Sains vs COVID-19 Sévère.

Pour créer ces Volcano-plots, nous avons calculé deux paramètres : le Fold-change (FC), qui reflète l'effet biologique, c'est-à-dire le degré de changement entre deux conditions, et la p-value, qui mesure l'importance statistique de cet effet. Les p-values ont été obtenues grâce au test t de Student, tandis que le FC a été calculé en prenant le rapport des moyennes de chaque condition.

Figure 11. Comparaison de l'expression génique entre différents groupes



Comme illustré dans la figure 11.A, entre les patients sains et ceux atteint de COVID-19, un total de 1591 gènes différentiellement exprimés a été obtenu avec 751 gènes sous-exprimés et 840 gènes sur-exprimés. Entre les patients sains et ceux atteint de COVID-19 Léger, 1244 sont différentiellement exprimés (556 sous-exprimés et 688 sur-exprimés). Entre les patients sains et ceux atteints de COVID-19 Sévère, 1591 sont différentiellement exprimés (751 sous-exprimés et 840 sur-exprimés).

Dans ces trois comparaisons, les gènes les plus différentiellement exprimés sont en grande partie similaires. Parmi eux, on retrouve notamment COA5, un gène codant pour une chaperonne impliquée dans l'organisation de la chaîne respiratoire mitochondriale.

En revanche, la comparaison entre les patients atteints de COVID-19 Léger et Sévère met en évidence des gènes différents. Parmi eux, ZNF680, un gène codant pour un site de liaison à un facteur de transcription impliqué dans la régulation de transcription de l'ADN.

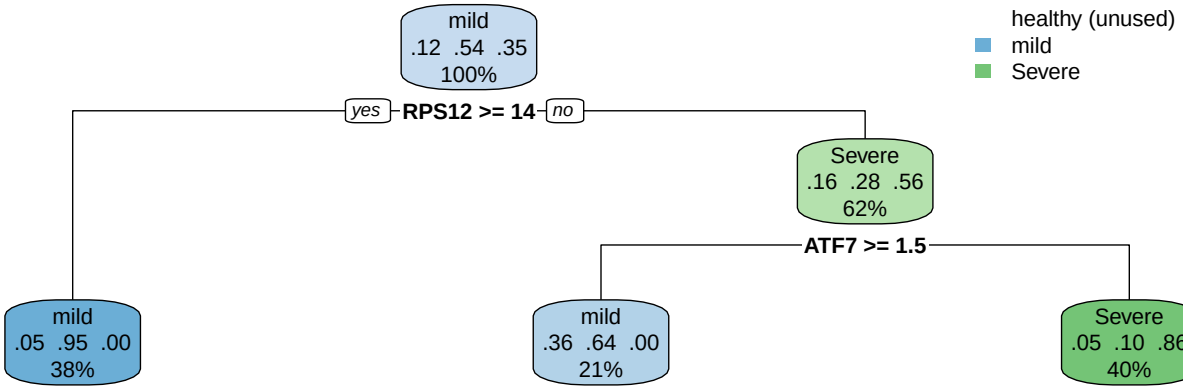
3.2 Analyse multivariée

Après avoir effectué l'analyse des gènes différentiellement exprimés, nous avons cherché à savoir s'il était possible de prédire la sévérité du COVID-19 en fonction de l'expression des gènes. Pour cela, nous avons réalisé une classification en machine learning.

3.2.1 Arbre de classification

Le modèle cherche è prédire la sévérité du COVID-19 en fonction des gènes exprimés. Il a été construit de la même manière que l'arbre de régression, à seule différence que ce dernier prédit une classe et non une valeur quantitative.

Figure 12. Arbre de classification



Comme l'illustre la figure 12, l'arbre de décision présente deux branches de division basées sur les gènes RSP12 et ATF7.

L'analyse de l'importance des variables (figure 13) révèle que les gène RPS12, ATF7 et HIF1AN ont le plus d'influence sur le modèle, avec des valeurs d'importance respectives de $9,53$, $8,14$ et $5,92$. Nous retrouvons également parmi les gènes importants, ZNF680, gène qui a été montré comme sur-exprimé dans l'analyse précédente (Figure 11.B).

Figure 13. Importance des variables

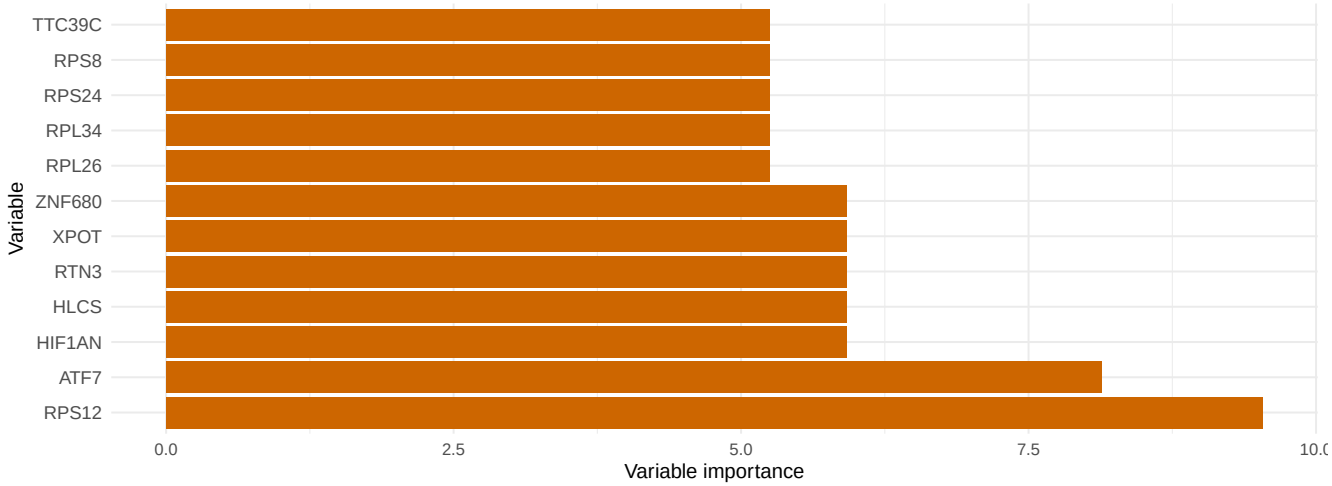
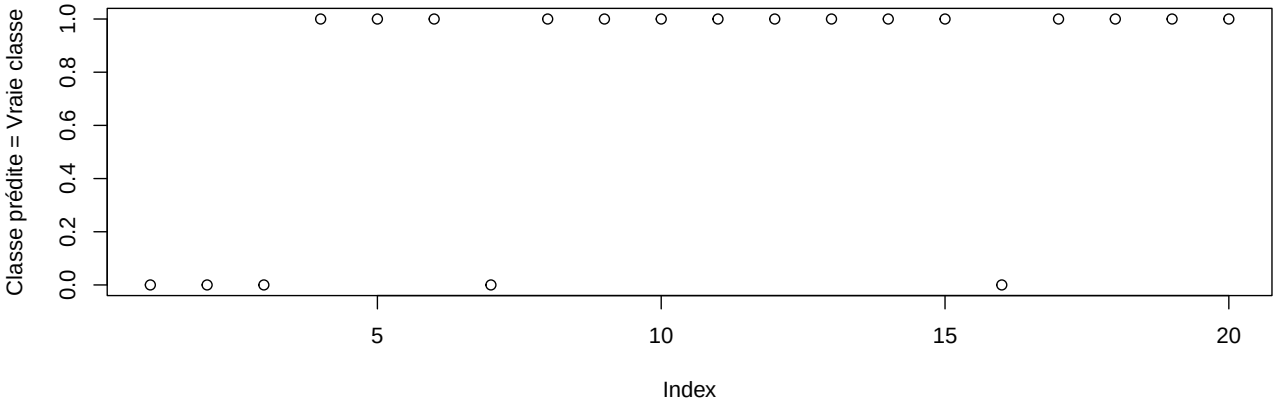


Figure 14. Précision du modèle



Le modèle a prédit correctement 75% des classes (Figure 14).

Néanmoins, afin de vérifier si le modèle n'a pas prédit au hasard, nous avons voulu comparer les proportions des classes obtenues par le modèle à celle qui aurait obtenuees par une classification aléatoire. Sachant que les proportions qui auraient été obtenues par une classification due au hasard sont de : $0,1$ $(2/10)$ d'individus sains, $0,55$ $(11/20)$ d'individus avec un COVID-19 Léger et de $0,35$ $(7/20)$ d'individus avec un COVID-19 Sévère.

Sachant aussi que le modèle a prédit 0 individus sains, une proportion de $0,55$ $(11/20)$ d'individus avec un COVID-19 Léger et $0,45$ $(9/20)$ d'individus avec un COVID-19 Sévère.

La probabilité de prédire au hasard étant égale à la somme des produits des proportions réelles et prédites, elle est de 0.46 pour un modèle au hasard, ce qui est loin des $0,75$ obtenu par notre modèle.

En conclusion, l'analyse des données RNA-seq a permis de mettre en lumière plusieurs gènes qui s'expriment différement selon le degré de sévérité de la COVID-19. La majorité de ces derniers sont

impliqués dans des processus essentiels au bon fonctionnel de la cellule dont la transcription de l'ADN ou encore la respiration cellulaire. Ainsi, un déreglement dans l'une de ces fonctions serait peut-être susceptible d'entrainer une détéerioration de l'état général des patients.

4 Conclusion

L'étude nous a permis d'analyser les facteurs impliqués dans la sévérité du COVID-19. L'analyse clinique a notamment mis en évidence plusieurs paramètres biologiques associés à la gravité de la maladie, tels que des marqueurs cellulaires témoignant d'inflammation et de lésions tissulaires, ainsi que des comorbidités comme l'hypertension, connues pour favoriser les troubles respiratoires. L'analyse transcriptomique quant à elle a permis de mettre en lumière des gènes dont une altération de l'expression pourrait également expliquer l'altération de l'état de santé du patient.

5 Références

Saji, Ryo (2021). COVID-19 Data base.csv. figshare. Dataset. https://doi.org/10.6084/m9.figshare.15059814.v1

Leem G & Shin E, GSE165461 Dataset, http://www.biomedical-web.com/covid19db/searchDetail?id=COVID000049