



HAL
open science

Prédiction du parcours de soins et du développement des enfants vulnérables

Lucas Anzelin

► **To cite this version:**

Lucas Anzelin. Prédiction du parcours de soins et du développement des enfants vulnérables. Santé publique et épidémiologie. Université Paris-Saclay, 2023. Français. NNT : 2023UPASR035 . tel-04426267

HAL Id: tel-04426267

<https://theses.hal.science/tel-04426267>

Submitted on 30 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Prédiction du parcours de soins et du développement des enfants vulnérables

Predicting the care trajectory and development of high-risk children

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°570 : santé publique (EDSP) Spécialité de
doctorat : Biostatistiques et data sciences
Graduate School : Santé publique. Référent : Faculté de médecine

Thèse préparée dans l'unité de recherche : **CESP (Université Paris-Saclay, UVSQ, Inserm)**,
sous la direction de **Anne THIÉBAUT**, Chargée de Recherche, INSERM
et la co-direction de **Matthieu HANF**, Directeur du Traitement de l'Information, SESAN

Thèse présentée et soutenue à Paris-Saclay, le 1^{er} décembre 2023, par

Lucas ANZELIN

Composition du jury

Membres du jury avec voix délibérative

Barbara HEUDE

Directrice de Recherche, Inserm,
Université Paris-Cité

Présidente & Rapporteur

Pierre-Yves ANCEL

PU-PH, Université Paris-Cité

Rapporteur & Examineur

Valérie GARES

Maîtresse de conférences,
Université de Rennes

Examinatrice

Grégoire REY

Directeur de France Cohortes,
Université Paris-Saclay

Examineur

Titre : Prédiction du parcours de soins et du développement des enfants vulnérables

Mots clés : Cohorte, Prématurité, Méthodes d'apprentissage

Résumé : Les enfants nés grands prématurés (avant 32 semaines d'aménorrhée) ou avec un très faible poids de naissance (moins de 1 500g) présentent un risque accru de développement altéré sur le long terme. Les directives internationales recommandent une surveillance médicale continue de ces enfants jusqu'à l'âge de sept ans pour détecter précocement les anomalies de développement et fournir des soins adaptés. Avec l'utilisation d'algorithmes statistiques avancés et afin de mieux anticiper les événements de santé futurs des patients, le secteur de la santé se transforme. Dans le suivi des enfants vulnérables, cette approche a le potentiel d'aider les professionnels à identifier les anomalies de développement et à optimiser les soins. En France, le suivi des enfants vulnérables est organisé via des réseaux de santé, coordonnés par les Agences Régionales de Santé. Depuis 2015, en Île-de-France (représentant 20% de la population française), les six réseaux de santé périnatale utilisent un système de dossier médical partagé, appelé HYGIE-SEV. Ce système de santé publique favorise le partage d'informations entre professionnels de santé et constitue une ressource pour la recherche scientifique à partir des données collectées.

Les objectifs de cette thèse étaient : 1) identifier les facteurs associés à un suivi non-optimal chez les enfants vulnérables suivis dans HYGIE-SEV, 2) élaborer des modèles prédictifs d'un suivi non-optimal, et 3) généraliser cette approche pour prédire les anomalies de développement. Les données comprenaient 14 557 enfants vulnérables suivis dans HYGIE-SEV entre novembre 2015 et août 2023, dont 12 344 (84,7%) grands prématurés et 12 007 (82,5%) de très faible poids de naissance. Les données incluaient un large éventail d'informations à la naissance telles que des caractéristiques socio-économiques, des caractéristiques cliniques sur la grossesse, l'accouchement et les soins néonataux, ainsi que des informations sur la gestion du suivi.

Concernant l'optimalité du suivi, 39,0% des 9 958 enfants éligibles avaient manqué leur visite à deux ans, et 62,6% des 3 881 enfants éligibles avaient manqué leur visite à cinq ans. Les facteurs de risque associés à la non-participation étaient similaires entre deux et cinq ans et comprenaient des facteurs liés à 1) de meilleures conditions médicales initiales telles qu'un âge gestationnel plus élevé, 2) des conditions socio-économiques plus défavorables et 3) des caractéristiques de la gestion du suivi, comme un médecin référent hospitalier. Une méthodologie robuste et reproductible a été mise en place pour évaluer plusieurs algorithmes appliqués à la prédiction de la non-participation aux visites à deux et cinq ans. Plusieurs algorithmes ont été testés, les Random Forests, les Support Vector Machine (SVM), XGBoost, les Multivariate Adaptive Regression Splines (MARS), la régression logistique avec et sans pénalisation, ainsi qu'un réseau de neurones à une couche cachée. Une méthode d'ensemble a également été utilisée. Parmi ces algorithmes, les Random Forests avaient les meilleures performances avec une aire sous la courbe ROC (AUC) de 77,8% à deux ans et de 76,6% à cinq ans pour la prédiction de la non-optimalité du suivi. En appliquant une approche similaire pour prédire le neurodéveloppement à la visite de deux ans, une performance AUC de 78,7% était obtenue.

Les résultats prometteurs observés dans la cohorte HYGIE-SEV soulignent l'importance de l'utilisation d'algorithmes prédictifs dans le suivi des enfants vulnérables et les soins de santé. Cette utilisation a le potentiel de favoriser le développement d'outils de e-santé visant à soutenir les professionnels de santé dans l'identification précoce des enfants à risque. En lien avec la médecine personnalisée cela pourrait permettre la délivrance de soins plus adaptés.

Title : Predicting the care trajectory and development of high-risk children

Keywords : Cohort, Preterm birth, Machine learning

Abstract : Infants born very preterm (before 32 weeks of gestation) or with a very low birth weight (under 1,500g) are at an elevated risk of long-term adverse outcomes. International guidelines recommend providing medical care and continuous monitoring for these high-risk children up to the age of seven to early detect developmental anomalies and provide appropriate care. The healthcare sector is undergoing a data-driven transformation with the use of advanced statistical algorithms to better anticipate future health outcomes in patients. Such approach in the follow-up of high-risk children has the potential to assist healthcare professionals in identifying early developmental issues and optimizing care and follow-up. In France, the monitoring of high-risk children is organized through specialized healthcare networks, coordinated by Regional Health Agencies. In the Île-de-France region (representing approximately 20% of the French population), the six specialized perinatal healthcare networks have used a shared electronic health record system, known as HYGIE-SEV, since 2015. This system promotes seamless information sharing among healthcare professionals, supports public health surveillance, and serves as a valuable resource for scientific research using collected data. The objectives of this thesis were to: 1) identify factors associated with suboptimal follow-up in high-risk children enrolled in the HYGIE-SEV program, 2) construct predictive models for suboptimal follow-up, and 3) generalize this approach to predict developmental anomalies. The dataset used in this thesis comprised data from 14,557 high-risk children included in HYGIE-SEV program from November 2015 to August 1, 2023, with 12,344 (84.7%) being born very preterm, and 12,007 (82.5%) having a very low birth weight. This dataset encompassed a wide range of information at birth, including socioeconomic characteristics, clinical details of pregnancy, delivery, and neonatal care, as well as information regarding follow-up management.

Concerning suboptimal follow-up, the data revealed that 39.0% of the 9,958 eligible children missed their two-year visit, and 62.6% of the 3,881 eligible children missed their five-year visit. Risk factors associated with non-attendance were consistent between the two- and five-year visits and included factors associated with 1) better initial medical conditions such as higher gestational age, 2) lower socio-economic conditions, and 3) characteristics of the follow-up management such as a hospital referring physician responsible for follow-up. A robust and reproducible methodology was then implemented to evaluate several algorithms applied to the prediction of non-attendance at the 2- and 5-years visits. This included data pre-processing, feature engineering, hyperparameter optimization, cross-validation, and managing imbalanced data. Multiple algorithms were tested, including Random Forests, Support Vector Machine (SVM), XGBoost, Multivariate Adaptive Regression Splines (MARS), logistic regression with and without penalization, and a single hidden layer neural network. An ensemble method was also employed. Among these algorithms, Random Forests demonstrated the best performance with an area under the ROC curve (AUC) of 77.8% at two years and 76.6% at five years for predicting non-attendance. Applying a similar approach to predict the optimality of neurodevelopment at the two-year visit yielded an AUC performance of 78.7%. The promising results observed in the HYGIE-SEV cohort underscore the importance of integrating predictive algorithms into the realms of high-risk children's follow-up and healthcare. This integration has the potential to pave the way for the development of eHealth tools aimed at supporting healthcare professionals in the early identification of at-risk children. Consequently, it could enable the delivery of more proactive and customized care, aligning with the principles of personalized medicine.



Prédiction du parcours de soins et du développement des enfants vulnérables

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°570 : santé publique (EDSP)
Spécialité de doctorat : santé publique - biostatistique
Graduate School : Santé publique Référent : Faculté de médecine

Thèse préparée dans l'unité de recherche : **CESP (Université Paris-Saclay, UVSQ, Inserm)**,
sous la direction de **Anne THIÉBAUT**, Chargée de Recherche, INSERM
et la co-direction de **Mathieu HANF**, Directeur du Traitement de l'Information, SESAN

Thèse présentée et soutenue à Villejuif, le 1^{er} décembre 2023, par

Lucas ANZELIN

Composition du jury

Membres du jury avec voix délibérative

Barbara HEUDE

Directrice de Recherche, Inserm,
Université Paris-Cité

Présidente & Rapporteur

Pierre-Yves ANCEL

PU-PH, Université Paris-Cité

Rapporteur & Examineur

Valérie GARES

Maîtresse de conférences,
Université de Rennes

Examinatrice

Grégoire REY

Directeur de France Cohortes,
Université Paris-Saclay

Examineur

REMERCIEMENTS

Tout d'abord je tiens à remercier l'ensemble des membres du jury de thèse de m'avoir fait l'honneur d'être rapporteurs et examinateurs de mon travail de thèse. Ainsi merci à Barbara Heude et Pierre-Yves Ancel d'avoir accepté d'être rapporteurs et merci à Valérie Garès et Grégoire Rey d'être examinateurs.

Je remercie également mes deux directeurs de thèse de m'avoir suivi au cours de ces quatre années écoulées. Tout d'abord merci à celui qui fut à l'initiative de ce travail, mon directeur Matthieu Hanf, qui m'a proposé d'effectuer cette thèse de science. Ce travail aura été une grande opportunité professionnelle et aura permis d'acquérir pléthore de compétences qui m'auront pour sûr fait grandir en tant que professionnel de santé publique. Ainsi, merci de m'avoir permis cette ouverture professionnelle et de m'avoir accompagné tout du long. J'en viens maintenant à remercier ma directrice de thèse Anne Thiébaud. Merci d'avoir été autant investie et d'avoir accordé autant de temps à l'ensemble de mes travaux. La méthodologie de travail à laquelle j'ai été confronté à ses côtés fut une aide précieuse pour la réalisation de cette thèse. Sa rigueur et son professionnalisme aura été très formateur. Cette initiation aux travaux de recherche raisonnera encore longtemps dans ma vie professionnelle, merci donc encore à mes deux directeurs de thèse de m'avoir permis cela.

Je tiens maintenant à remercier SESAN et sa directrice Naïma Mezaour, de m'avoir accordé le temps et les moyens nécessaires à la réalisation de cette thèse. Merci également à Pierre Boiron, ancien directeur, de m'avoir fait confiance ce 3 décembre 2018 pour décrocher ce premier travail en sortie d'étude. Cinq ans plus tard, cette thèse fut l'aboutissement du plus long des riches projets auxquels j'ai pu participé au sein de SESAN. Merci également à tous les collègues que j'ai côtoyés depuis mon arrivée d'avoir permis cette bonne ambiance de travail facilitant ainsi la réalisation de mes travaux. Je tiens à remercier spécialement les stagiaires ayant directement travaillé sur les sujets en lien avec cette thèse dont leur apport fut une aide précieuse. Merci donc à François et Sara pour leur contribution au présent travail. Merci également à Brice qui fut bien plus que simple collègue, Aurélie avec qui nous avons tout connu chez SESAN, Thomas I. et Joana mes collègues de périnatalité, Alper, Daouda, Maxime, Charline, Saif, mes collègues Data Scientist, sans oublier Fadhila, merci également à mes autres collègues du « Traitement de l'Information » Nabia, Mathias, Robin, Thomas G.,

Julien M, Oscar, Imane et Karen avec qui je partage l'équipe et « *l'open space* ». Merci à tous les autres également Valentin, Agathe, Mimie, Samuel, Clément, Julien P... pour tous ces moments de partage pendant et en dehors des horaires de travail.

Un grand merci également à Pascale Tubert-Bitter, de m'avoir permis de travailler dans son équipe de Biostatistique en grande dimension, au sein du Centre de recherche en Epidémiologie et en Santé des Populations (CESP). Son accueil chaleureux, sa bienveillance et ses conseils avisés m'auront également permis de me sentir à l'aise dans ce deuxième environnement de travail. De même, merci à Sylvie et Ismaïl pour leurs conseils et pour tous les moments d'échange à Villejuif. Merci aussi à tous les doctorants avec qui nous avons pu échanger sur nos travaux. De façon non exhaustive merci donc à Etienne, Matthieu, Liliane, Hong, Emeline.... Encore une fois merci à eux.

Merci à tous les investigateurs, médecins et coordinateurs de HYGIE-SEV sans qui je n'aurais pas pu effectuer ces travaux. Ainsi, merci à l'Agence Régionale de Santé d'Île-de-France (ARS-IdF) et à l'ensemble des réseaux de périnatalité de la région. Merci aussi à tous les membres du conseil scientifique de HYGIE-SEV. Merci à Laurence Desplanques, qui fut à l'origine de sa création. Enfin un merci spécial à Michèle Granier, directrice du conseil scientifique et à Laurence Leloup avec qui c'est toujours un plaisir de travailler.

Cette thèse ayant bénéficié d'une Convention Industrielle de Formation par la Recherche (CIFRe), j'en profite pour remercier le ministère de la recherche et de l'enseignement supérieur et en particulier l'Association Nationale de la Recherche Technologique (ANRT) pour leur confiance et leur investissement financier dans ses travaux. Par ailleurs je remercie également toute l'équipe de l'école doctorale de Santé Publique.

Je remercie mes amis de toujours qui, lors des moments de détente furent une grande bouffée d'oxygène au cours de ces quatre années. Merci à Baptiste et Nono, merci à Julien, Manon et mon filleul Jojo, merci à Thomas le canadien, à Kylian et à tous les autres, avec qui, depuis plus de dix ans partagent tous ces moments de convivialité. Merci malgré tout à Lison pour son accompagnement et ses conseils durant la première partie de cette thèse. Merci à ma famille qui m'aura accompagné depuis le début de la thèse (en réalité depuis bien avant) jusqu'à l'aboutissement de celle-ci. Merci à mes parents, mon frère et mes grands-parents pour ce soutien sans faille envers tous mes projets professionnels (et personnels) même si

ceux-ci doivent s'exécuter dans la capitale. Je dois une grande partie de la réalisation de cette thèse à leur présence à mes côtés depuis toujours et au respect de chacun de mes choix. Enfin, merci surtout à celle qui fut à mes côtés depuis ces derniers mois et notamment dans la rédaction de ce manuscrit. Merci donc à Charlotte pour m'avoir apporté ce second souffle dont j'avais besoin dans cette période difficile, merci pour sa compréhension et son dévouement.

Enfin je tenais à remercier très sincèrement les enfants du programme HYGIE-SEV ainsi que leur entourage, qui, en étant les principaux concernés par ce lourd suivi, se retrouvent dans chaque ligne de l'écriture de cette thèse. Pour ne pas oublier que derrière chaque chiffre présenté ici se trouvent des familles confrontées à un quotidien difficile, je leur dédie cette thèse.

PRODUCTIONS EN COURS ET REALISEES

Article soumis :

Lucas Anzelin, Matthieu Hanf, Laurence Desplanques, Michèle Granier, Laurence Leloup, Alexandre Lapillonne, Sylvie Escolano, Pascale Tubert-Bitter, Anne C.M. Thiébaud.

Factors associated with suboptimal follow-up of very preterm infants at two and five years in a large population-based French healthcare network.

Article en cours de rédaction :

Cohort profile: Longitudinal study of infants born very preterm or with a very low birth weight in the Ile de France region of France (HYGIE-SEV cohort).

Présentation orale :

International Biometric Conferences, IBC, Riga, juillet 2023 :

Lucas Anzelin, Matthieu Hanf, Ismaïl Ahmed, Pascale Tubert-Bitter, Laurence Desplanques, Michèle Granier, Anne C.M. Thiébaud.

Comparison of algorithms to predict loss-to follow-up at two years and five years of age among very preterm children in the Île-de-France region.

Communications affichées :

Statistics and Biopharmacy, SnB, Paris, septembre 2023 :

Lucas Anzelin, Matthieu Hanf, Ismaïl Ahmed, Pascale Tubert-Bitter, Laurence Desplanques, Michèle Granier, Anne C.M. Thiébaud.

Comparison of algorithms to predict loss-to follow-up at two years and five years of age among very preterm children in the Île-de-France region.

International Society for Clinical Biostatistics, ISCB, Milan, août 2023 :

Lucas Anzelin, Anne C.M. Thiébaud, Laurence Desplanques, Michèle Granier, Laurence Leloup, Véronique Pierrat, Alexandre Lapillonne, Pascale Tubert-Bitter, Ismaïl Ahmed, Matthieu Hanf

Prediction of neurodevelopment at two years of age among very preterm infants in the Île-de-France region.

TABLE DES MATIERES

REMERCIEMENTS.....	4
PRODUCTIONS EN COURS ET REALISEES	7
TABLE DES MATIERES.....	9
LISTE DES TABLEAUX.....	15
LISTE DES FIGURES.....	18
LISTE DES ABBREVIATIONS.....	19
CHAPITRE 1 – INTRODUCTION GENERALE.....	21
1. Suivi des enfants vulnérables.....	21
1.1. Mortalité chez les enfants de moins de 5 ans	21
1.2. Incidence et facteurs de risque de prématurité.....	24
1.3. Développement des enfants nés prématurément et autres enfants vulnérables ...	26
1.4. Importance de la mise en place d'un suivi.....	27
1.4.1. Prise en charge des enfants vulnérables	27
1.4.2. Problématiques liées au suivi des enfants vulnérables.....	29
2. Médecine personnalisée et suivi des enfants vulnérables	31
3. Algorithmes de prédiction.....	36
3.1. Définition générale et contexte d'utilisation.....	36
3.2. Présentation des algorithmes de prédiction	37
3.2.1. Régression logistique.....	38
3.2.2. Régression LASSO.....	39
3.2.3. Régression multivariée par splines adaptatives (MARS).....	39
3.2.4. Random Forests	40
3.2.5. Extreme Gradient Boosting (XGBoost)	42
3.2.6. Support Vector Machine (SVM)	44

3.2.7.	Réseaux de neurones	46
3.2.8.	Méthode d'ensemble	48
3.3.	Optimisation d'un algorithme de prédiction	49
3.3.1.	Pré-traitement des données.....	49
3.3.2.	Optimisation des hyperparamètres.....	52
3.3.3.	Validation croisée	54
3.4.	Evaluation de la performance	54
4.	Objectifs de recherche	58
4.1.	Comprendre les problématiques liées au suivi des enfants vulnérables.....	58
4.2.	Utiliser des techniques innovantes pour améliorer la prise en charge des enfants vulnérables.....	58
4.3.	Renforcer le suivi individualisé	58
CHAPITRE 2 – DONNEES : LE DISPOSITIF HYGIE-SEV.....		61
1.	Présentation de HYGIE-SEV	61
1.1.	Contexte et justification du programme.....	61
1.2.	Objectifs du programme.....	64
2.	Protocole de suivi des enfants vulnérables.....	64
2.1.	Critères d'inclusion du programme	64
2.2.	Suivi des enfants inclus.....	65
2.3.	Plateforme informatique.....	66
3.	Données collectées	68
3.1.	Caractéristiques socio-économiques	70
3.2.	Caractéristiques de la grossesse et du nouveau-né.....	70
3.3.	Caractéristiques de la prise en charge médicale	71
4.	Etude descriptive : population et méthodes	72
5.	Résultats descriptifs : caractéristiques à l'inclusion.....	72

5.1.	Caractéristiques socio-économiques	73
5.2.	Caractéristiques de la grossesse	76
5.3.	Caractéristiques néonatales.....	78
5.1.	Caractéristiques de prise en charge médicale	79
6.	Forces et limites du programme.....	81
CHAPITRE 3 – OPTIMALITE DU SUIVI : ANALYSE DES FACTEURS ASSOCIES		85
1.	Introduction	85
1.1.	Problématique générale.....	85
1.2.	Problématiques liées au suivi en Île-de-France.....	86
2.	Méthodes.....	88
2.1.	Population incluse	88
2.2.	Variables d'exposition.....	90
2.3.	Définition du critère de jugement.....	91
2.4.	Analyses statistiques.....	92
3.	Résultats	93
3.1.	Analyses descriptives.....	93
3.2.	Résultats à 2 ans.....	99
3.3.	Résultats à 5 ans.....	105
3.4.	Analyse de sensibilité.....	105
4.	Discussion.....	110
4.1.	Principaux résultats.....	110
4.2.	Comparaison avec la littérature	110
4.3.	Forces et limites de l'étude	112
4.4.	Perspectives	113
CHAPITRE 4 – OPTIMALITE DU SUIVI : ANALYSES PREDICTIVES.....		115
1.	Introduction	115

2.	Méthodologie.....	117
2.1.	Population d'étude.....	117
2.2.	Définition de la variable de sortie	117
2.3.	Variables prédictives.....	117
2.4.	Modèles prédictifs.....	118
2.5.	Pré-traitement des données.....	120
2.6.	Grilles de recherche	121
2.7.	Validation croisée	122
2.8.	Evaluation	123
2.9.	Mesure d'importance des variables prédictives.....	124
3.	Résultats.....	124
3.1.	Sélection de la population	124
3.2.	Résultats à deux ans	126
3.2.1.	Optimisation des hyperparamètres.....	126
3.2.2.	Optimisation de la méthode d'ensemble.....	126
3.2.3.	Evaluation sur les données de test	127
3.2.4.	Importance des variables prédictives	128
3.3.	Résultats à cinq ans	129
3.3.1.	Optimisation des hyperparamètres.....	129
3.3.2.	Optimisation de la méthode d'ensemble.....	130
3.3.3.	Evaluation sur les données de test	131
3.3.4.	Importance des variables prédictives	131
3.4.	Résultats du modèle incrémental.....	132
4.	Discussion.....	133
4.1.	Synthèse des résultats	133
4.2.	Comparaison avec la littérature	134

4.3.	Forces et limites de l'étude	135
4.4.	Perspectives	136
CHAPITRE 5 –	GENERALISATION DES ALGORITHMES : OPTIMALITE DU	
NEURODEVELOPPEMENT		139
1.	Introduction	139
2.	Méthodologie.....	140
2.1.	Population d'étude.....	140
2.2.	Définition de la variable de sortie	140
2.3.	Modèle prédictif et optimisation.....	141
2.4.	Analyses statistiques.....	142
3.	Résultats.....	142
4.	Discussion.....	149
4.1.	Synthèse des résultats	149
4.2.	Comparaison avec la littérature	150
4.3.	Comparaison avec l'étude des perdus de vue.....	150
4.4.	Forces et limites de l'étude	151
4.5.	Perspectives	152
CHAPITRE 6 –	DISCUSSION GENERALE.....	155
1.	Synthèse des résultats obtenus.....	155
1.1.	Le système d'information HYGIE-SEV.....	155
1.2.	Complexité de la question des perdus de vue	157
1.3.	Des pistes d'amélioration pour le suivi personnalisé.....	159
2.	Perspectives	160
2.1.	Enrichissement des données pour compléter le parcours de soins.....	160
2.2.	Perspectives d'amélioration des algorithmes	163

2.3. Généralisation des algorithmes et création d'un outil à destination des professionnels.....	164
3. Conclusion.....	165
BIBLIOGRAPHIE.....	166
ANNEXES.....	189
Annexe 1 : Tableau de synthèse des études publiées sur les enfants vulnérables.....	189
Annexe 2 : Extraction du manuel d'exploitation des données de HYGIE-SEV fourni par l'industriel.....	190
Annexe 3 : Scores utilisés dans le programme HYGIE-SEV	191
Annexe 4 : Analyses complémentaires.....	223

LISTE DES TABLEAUX

Tableau 1 : Définition des différents types de mortalité périnatale communément utilisés par les Nations Unies (3).....	22
Tableau 2 : Définition de la prématurité.....	24
Tableau 3 : Hyperparamètres présents dans un SVM.....	45
Tableau 4 : Hyperparamètres présents dans un réseau de neurones.....	48
Tableau 5 : Données collectées dans le programme HYGIE-SEV.....	68
Tableau 6 : Caractéristiques socio-économiques des enfants à l'inclusion (N=14 557), cohorte HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023.....	74
Tableau 7 : Caractéristiques de la grossesse et de l'accouchement à l'inclusion (N=14 557), cohorte HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023.....	77
Tableau 8 : Caractéristiques néonatale à l'inclusion (N=14 557), cohorte HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023.....	80
Tableau 9 : Caractéristiques de la prise en charge à l'inclusion (N=14 557), cohorte HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023.....	81
Tableau 10 : Caractéristiques des patients à l'inclusion (N=14 483), à la visite à deux ans (N=9 958) et à la visite à cinq ans (N=3 898), cohorte HYGIE-SEV, 1er novembre 2015 - 31 octobre 2022.....	94
Tableau 11 : Facteurs associés à l'absence de visite à deux ans dans les régressions logistiques univariées et multivariées après imputation multiple, cohorte HYGIE-SEV, 1er novembre 2015-31 octobre 2022 (N=9 958).....	101
Tableau 12 : Facteurs associés à l'absence de visite à cinq ans dans les régressions logistiques univariées et multivariées après imputation multiple, cohorte HYGIE-SEV, 1er novembre 2015-31 octobre 2022 (N=3 898).....	106
Tableau 13 : Résultats de l'optimisation des hyperparamètres dans les modèles prédictifs de l'absence de visite à deux ans, cohorte HYGIE-SEV, échantillon d'entraînement (N = 7 201).....	126
Tableau 14 : Résultats de la méthode d'ensemble pour la prédiction de l'absence de visite à deux ans, cohorte HYGIE-SEV, échantillon d'entraînement (N = 7 201).....	127
Tableau 15 : Evaluation de la performance des modèles prédictifs de l'absence de visite à deux ans, cohorte HYGIE-SEV, échantillon test (N = 1 801).....	127

Tableau 16 : Extraction des 20 variables avec la plus grande mesure d'importance dans le modèle des <i>Random Forests</i> pour la prédiction de l'absence de visite à deux ans, cohorte HYGIE-SEV, échantillon test (N = 1 801).....	129
Tableau 17 : Résultats de l'optimisation des hyperparamètres dans les modèles prédictifs de l'absence de visite à cinq ans, cohorte HYGIE-SEV, échantillon d'entraînement (N = 2 586)	130
Tableau 18 : Résultats de la méthode d'ensemble pour la prédiction de l'absence de visite à cinq ans, cohorte HYGIE-SEV, échantillon d'entraînement (N = 2 586).....	131
Tableau 19 : Evaluation de la performance des modèles prédictifs de l'absence de visite à cinq ans à cinq ans, cohorte HYGIE-SEV, échantillon test (N = 648).....	131
Tableau 20 : Extraction des 20 variables avec la plus grande mesure d'importance dans le modèle des <i>Random Forests</i> pour la prédiction de l'absence de visite à cinq ans, cohorte HYGIE-SEV, échantillon test (N = 648)	132
Tableau 21 : Principales caractéristiques des enfants sur les critères d'optimalité du neurodéveloppement à deux ans (N = 6 546)	144
Tableau 22 : Principales caractéristiques des enfants selon l'optimalité du neurodéveloppement à deux ans (N = 6 546)	146
Tableau 23 : Résultats de l'optimisation des hyperparamètres dans le modèle de prédiction du neurodéveloppement à deux ans, cohorte HYGIE-SEV, échantillon d'entraînement (N = 5 236).....	147
Tableau 24 : Résultats des métriques des <i>Random Forests</i> sur les données d'entraînement.	148
Tableau 25 : Evaluation de la performance du modèle de <i>Random Forests</i> pour la prédiction du neurodéveloppement à deux ans, cohorte HYGIE-SEV, échantillon de test (N = 1 310) ..	148
Tableau 26 : Extraction des 20 variables avec la plus grande mesure d'importance dans le modèle des <i>Random Forests</i> pour la prédiction de d'un neurodéveloppement non-optimal, cohorte HYGIE-SEV, échantillon test (N = 1 310).....	149
Tableau S1 : Synthèse des principales cohortes d'enfants vulnérables dans le monde	189
Tableau S2 : Description des scores à 4 mois	191
Tableau S3 : Description des scores à 1 an.....	192
Tableau S4 : Description des scores à 2 ans.....	195
Tableau S5 : Description des scores à 3 ans.....	199
Tableau S6 : Description des scores à 4 ans.....	203
Tableau S7 : Description des scores à 5 ans.....	208

Tableau S8 : Description des scores à 6 ans.....	213
Tableau S9 : Description des scores à 7 ans.....	218
Tableau S10 : Facteurs associés à l'absence de visite à deux ans dans les régressions logistiques univariées et multivariées avant imputation multiple, cohorte HYGIE-SEV, 1er novembre 2015-31 octobre 2022 (N=9 958).....	223
Tableau S11 : Facteurs associés à l'absence de visite à cinq ans dans les régressions logistiques univariées et multivariées avant imputation multiple, cohorte HYGIE-SEV, 1er novembre 2015-31 octobre 2022 (N=3 898).....	228
Tableau S12 : Facteurs associés à la perte de vue à deux ans (définition plus stricte) dans les régressions logistiques univariées et multivariées après imputation multiple, cohorte HYGIE-SEV, 1er novembre 2015-31 octobre 2022 (N=3 898).....	233

LISTE DES FIGURES

Figure 1 : Taux de mortalité des moins de 5 ans par pays, 2015 (4)	23
Figure 2 : Taux de prématurité, 2019 (14)	25
Figure 3 : Typologie des informations individuelles utilisées en médecin personnalisée.....	32
Figure 4 : Classification fondamentale des algorithmes de prédiction	37
Figure 5 : Exemple d'un arbre de classification d'une variable à prédire binaire et de trois variables prédictives binaires	41
Figure 6 : Schéma simplifié de l'hyperplan linéaire d'un modèle de SVM.....	44
Figure 7 : Exemple d'un réseau de neurone à trois variables d'entrée, une couche cachée de cinq neurones et une variable de sortie binaire	47
Figure 8 : Utilisation des combinaisons pour la grille de recherche.....	53
Figure 9 : Exemple d'une matrice de confusion à deux entrées	55
Figure 10 : Exemple d'une courbe ROC.....	57
Figure 11 : Répartition des réseaux de santé en périnatalité en Île-de-France.....	62
Figure 12 : Répartition des lieux d'inclusion (maternités) et de suivi (médecins pilotes) en perspective du niveau urbain dans HYGIE-SEV, région Île-de-France, 2015-2022.....	65
Figure 13 : Exemple d'une timeline dans HYGIE-SEV.....	66
Figure 14 : Sélection de la population d'étude pour la description des données HYGIE-SEV SA, semaines d'aménorrhée.....	73
Figure 15 : Evolution des proportions de visites réalisées (taux de suivi, en vert) et non réalisées (taux d'attrition, en rouge) dans le programme HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023	87
Figure 16 : Diagramme de flux des enfants inclus dans les analyses à deux et cinq ans, cohorte HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023.....	89
Figure 17 : Sélection de la population éligible aux analyses prédictives de l'optimalité du suivi	125
Figure 18 : Sélection de la population éligible aux analyses prédictives du neurodéveloppement, cohorte HYGIE-SEV	143
Figure S1 : Exemple de schéma de données du manuel d'exploitation fourni par l'industriel pour comprendre l'architecture de données de HYGIE-SEV.....	190

LISTE DES ABBREVIATIONS

ADASYN – *ADaptive SYNthetic*

AG – Age Gestationnel

AME – Aide Médicale de l'État

ANRT – Association Nationale de la Recherche et de la Technologie

ARS – Agence Régionale de Santé

ASQ – *Age State Questionnaire*

AUC – *Area Under the Curve*

CIFRe – Convention Industrielle de Formation par la Recherche

CMU – Couverture Maladie Universelle

FN – Faux Négatif

FP – Faux Positif

IA – Intelligence Artificielle

IC – Intervalle de Confiance

IDF – Île-de-France

IM – Imputation Multiple

INSEE – Institut National de la Statistique et des Etudes Economiques

k-NN – *k-Nearest Neighbors*

LASSO – *Least Absolute Shrinkage and Selection Operator*

MARS – *Multivariate Adaptive Regression Splines*

MYPAs – Maternité en Yvelines et Périnatalité Active

NEF – Naître dans l'Est Francilien

OMS – Organisation Mondiale de la Santé

OR – *Odds Ratio*

PMA – Procréation Médicalement Assistée

RBF – *Radial Basis Function*

ROC – *Receiver Operating Characteristic*

RPSOF – Réseau Pédiatrique Sud et Ouest Francilien

RPVM – Réseau Périnatal du Val-de-Marne

RPVO – Réseau Périnatal du Val d'Oise

RSP – Réseau de Santé en Périnatalité
RSPP – Réseau de Santé Périnatal Parisien
SA – Semaines d'Aménorrhée
SEV – Suivi des Enfants Vulnérables
SI – Système d'Information
SNDS – Système National des Données de Santé
SVM – *Support Vector Machine*
VPN – Valeur Prédictive Négative
VPP – Valeur Prédictive Positive
VIF – *Variance Inflation Factor*
VN – Vrai Négatif
VP – Vrai Positif
VRS – Virus Respiratoire Syncytial
XGBoost – *eXtreme Gradient Boosting*

CHAPITRE 1 –

INTRODUCTION GENERALE

Dans ce chapitre introductif, nous détaillons le contexte dans lequel se situe ce travail de thèse, à savoir le suivi des enfants vulnérables (Section 1). Nous présentons ensuite les enjeux que pose ce suivi dans le contexte de la médecine personnalisée (Section 2), puis les outils statistiques permettant de répondre à ces enjeux, à savoir les algorithmes de prédiction (Section 3). Le contexte étant ainsi posé, ce chapitre se termine par l'énoncé des objectifs du travail de thèse (Section 4).

1. Suivi des enfants vulnérables

1.1. Mortalité chez les enfants de moins de 5 ans

Parmi les huit objectifs du millénaire pour le développement (MDG pour *Millennium Development Goals*) des Nations Unies (1), figure celui de réduire de deux tiers le nombre de décès évitables chez les enfants de moins de cinq ans entre 1990 et 2015, faisant de la santé périnatale et pédiatrique un enjeu majeur de développement humain (2). En santé périnatale, plusieurs types de mortalité sont définies : mortinatalité, mortalité néonatale et mortalité infantile (tableau 1). Les taux de mortinatalité, de mortalité néonatale, de mortalité infantile et de mortalité des enfants de moins de 5 ans sont calculés par le rapport entre le nombre d'enfants décédés dans la classe d'âge ou la période périnatale concernée et le nombre de naissances vivantes.

Tableau 1 : Définition des différents types de mortalité périnatale communément utilisés par les Nations Unies (3)

Mortalité	Définition
Mortinatalité	Nombre d'enfants mort-nés
Mortalité néonatale	Nombre d'enfants nés vivants et décédés avant 27 jours
- précoce	Entre 0 et 6 jours
- tardive	Entre 7 et 27 jours
Mortalité infantile	Nombre d'enfants nés vivants et décédés entre 0 et 1 an
Mortalité avant 5 ans	Nombre d'enfants nés vivants et décédés entre 0 et 5 ans

A la suite d'une initiative internationale ayant pour objectifs d'améliorer l'espérance de vie et de diminuer la mortalité dans chaque pays, les Nations Unies ont impulsé un programme mondial de collecte et analyse de données de mortalité et handicap intitulé « *Global Burden of Disease* » (GBD) et qui s'intéresse entre autres à la santé périnatale (4). Les chiffres rapportés montrent que, depuis plusieurs décennies, les taux de mortalité chez le jeune enfant sont en constante diminution dans le monde. Ainsi, le nombre de mort-nés pour 1 000 naissances vivantes est passé de plus de 30 dans les années 1980 à moins de 15 en 2015. De même, en ce qui concerne la mortalité néonatale précoce (0-6 jours), néonatale tardive (7-27 jours) et post-néonatale (28 jours-1 an), les taux sont passés respectivement d'environ 25, 10 et 35 pour 1 000, dans les années 1980 à environ 15, 4 et 13 pour 1 000 en 2015. Enfin la mortalité des enfants de moins de 5 ans est passée de 110 pour 1 000 naissances vivantes dans les années 1980 à moins de 50 décès pour 1 000 naissances vivantes dans les années 2010. Soit une diminution proche de 50 % pour tous ces indicateurs (4).

Malgré cette diminution sans précédent, les Nations Unies ne souhaitent pas s'arrêter là et réaffirment aujourd'hui un nouvel objectif à l'horizon 2030 dans leur nouveau programme plus global (17 objectifs) de développement durable (SDG pour *Sustainable Development Goals*) : une cible de moins de 12 décès néonataux pour 1 000 naissances vivantes dans tous les pays et de moins de 25 décès avant 5 ans pour 1 000 naissances vivantes (5). Cette volonté continue d'améliorer les indicateurs de mortalité chez les nouveau-nés et les enfants à l'échelle mondiale montre, d'une part, l'importance donnée à la santé périnatale et pédiatrique et, d'autre part, la marge de progression qu'il reste à accomplir dans ce domaine.

Bien que l'on observe une diminution globale de la mortalité périnatale et infantile sur l'ensemble des pays, l'étude du GBD (4) montre une disparité selon les zones géographiques.

On constate notamment une plus grande mortalité chez les enfants de moins de cinq ans dans les pays d’Afrique sub-saharienne et en Asie du Sud-Ouest. Inversement, les taux les plus favorables sont observés dans les pays d’Europe de l’Ouest et dans les pays d’Asie du Pacifique à haut revenu (d’après la banque mondiale un pays à haut revenu est défini par un revenu national brut par habitant de plus de 13 205 dollars américains (6)). La figure 1 illustre ces inégalités de mortalité chez les enfants de moins de cinq ans en 2015.

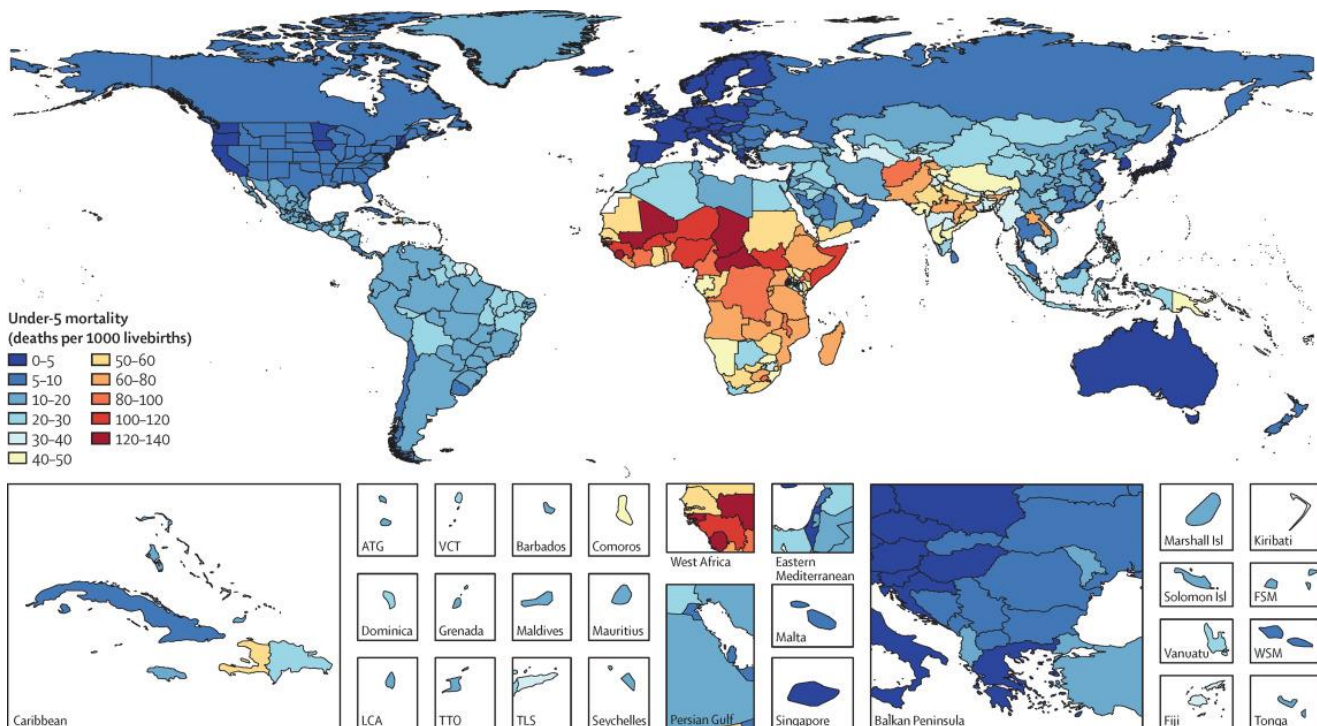


Figure 1 : Taux de mortalité des moins de 5 ans par pays, 2015 (4)

En France, on comptait en 2015 3,3 enfants mort-nés pour 1 000 naissances vivantes et respectivement 1,5, 1,2 et 3,3 décès néonataux précoces, néonataux tardifs et d’enfants de moins de 5 ans pour 1 000 naissances vivantes (4). Bien qu’on note une diminution de la mortalité depuis les années 1980, cette diminution est moins importante comparée à d’autres pays à haut revenu et a même tendance à se stabiliser depuis les années 2010 (3,7,8). Une étude récente a même mis en évidence une augmentation significative de la mortalité infantile en France entre 2012 et 2020 (9). Pour surveiller voire inverser cette tendance préoccupante en France, une collaboration étroite entre les professionnels de santé (médecins, sages-femmes, etc.), les chercheurs spécialisés en santé pédiatrique et périnatale,

ainsi que les institutions régionales et nationales de santé est devenue une nécessité incontournable.

1.2. Incidence et facteurs de risque de prématurité

Parmi les principales causes de mortalité, la naissance prématurée occupe une place prépondérante (4,10,11). Selon l'Organisation mondiale de la santé (OMS), la prématurité est définie comme la naissance d'un enfant avec un âge gestationnel (AG) inférieur à 37 semaines d'aménorrhée (SA) (12). Toujours d'après l'OMS elle est subdivisée en trois types : la prématurité modérée à tardive, la grande prématurité ou l'extrême prématurité (tableau 2).

Tableau 2 : Définition de la prématurité

Type de prématurité	Définition
Modérée à tardive	Naissance entre 32 SA et moins de 37 SA
Grande	Naissance entre 28 SA et moins de 32 SA
Extrême	Naissance de moins de à 28 SA

Le taux de prématurité est le rapport entre le nombre de naissances prématurées et le nombre de naissances vivantes. Une analyse systématique publiée en 2012 estime le taux de prématurité mondial à 11,1 % en 2012 (13). Cette étude révèle également des variations de taux de prématurité selon les pays, atteignant environ 18 % en Afrique et 5 % en Europe. Une étude plus récente en 2019 rapporte un taux de naissances prématurées de 10,6 % (14). Cette étude met également en évidence des disparités régionales, avec un taux de prématurité plus élevé en Afrique et un taux plus bas en Europe. Cependant, en Europe, par rapport à l'étude de 2012, on observe une augmentation du nombre de naissances prématurées, passant de 5 % en 2012 à 8,7 % en 2019. En Afrique, ce taux est passé de 18 % en 2012 à 12 % en 2019. La figure 2 illustre les disparités territoriales du taux de naissances prématurées à travers le monde.

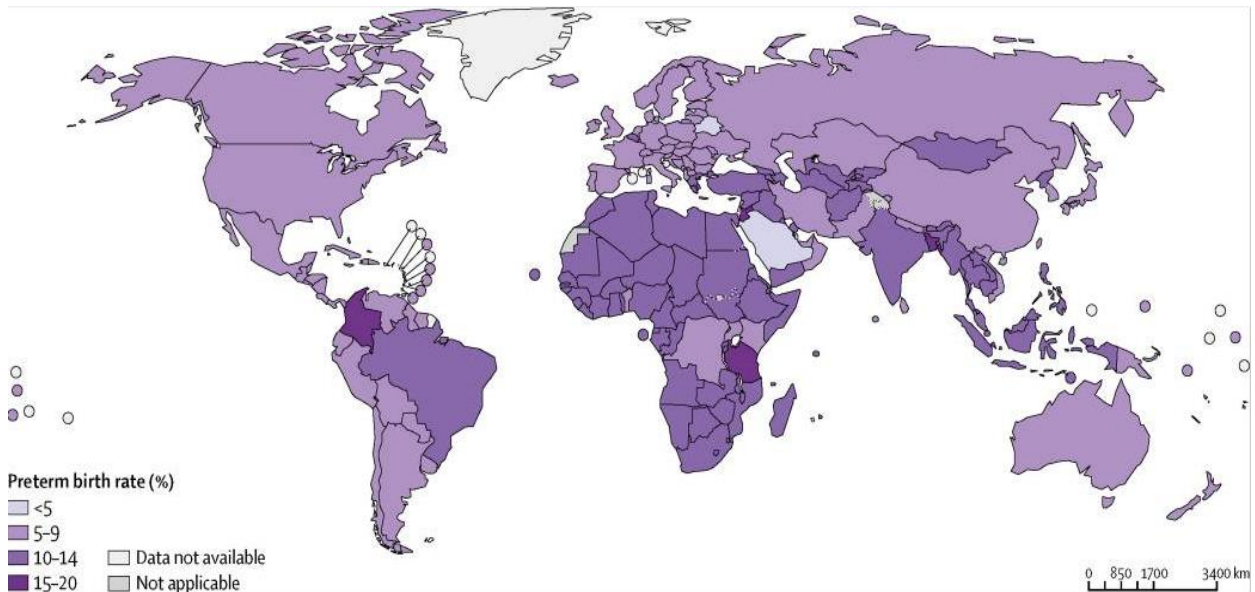


Figure 2 : Taux de prématurité, 2019 (14)

Comme précisé dans la méta-analyse de Chawanpaiboon et al. (14), il faut être prudent dans l'interprétation des tendances à l'augmentation ou à la diminution des taux de naissances prématurées car les méthodes de recueil et les définitions peuvent varier selon les pays et selon les années.

En France, comme dans la plupart des pays à revenu élevé, la naissance prématurée est un problème majeur, étant considérée comme la principale cause de mortalité néonatale et la deuxième cause de décès chez les enfants de moins de 5 ans, juste après les malformations congénitales (10,13). Il existe plusieurs facteurs de risque identifiés qui peuvent contribuer à la survenue d'une naissance prématurée.

Des facteurs socio-économiques ont été identifiés comme pouvant moduler le risque de prématurité (15-24). Ces facteurs socio-économiques incluent le jeune âge de la mère (inférieur à 18 ans) (21,23,25), le faible niveau d'éducation de la mère (niveau primaire ou non scolarisé) par rapport à un plus haut niveau (études supérieures) (24,26), le statut socio-économique des parents (ressources financières instables, absence de profession) (15,17-20), ainsi que la difficulté d'accès aux soins prénatals due à une vulnérabilité sociale (personne migrante, en situation de handicap ou ayant des problèmes psychologiques) (16,20).

Les complications pendant la grossesse et les antécédents maternels semblent également jouer un rôle important dans la survenue de la prématurité (11,15,17,27,28). Certaines conditions médicales telles que le diabète gestationnel (29,30), l'hypertension artérielle (31,32), les infections intra-utérines (33), les anomalies du col de l'utérus (34,35) et les antécédents de naissances prématurées (36–38) ont été associées à une augmentation du risque de donner naissance prématurément.

Enfin, les conditions environnementales du lieu de vie de la mère pendant la grossesse peuvent également jouer un rôle dans la survenue de la prématurité. Par exemple, le tabagisme pendant la grossesse a été fortement associé à un risque accru de prématurité (39). La pollution urbaine, en particulier la pollution de l'air, a également été identifiée comme un facteur de risque potentiel (40). De plus, certaines études suggèrent une corrélation entre les températures élevées et un risque accru d'accouchement prématuré (41).

1.3. Développement des enfants nés prématurément et autres enfants vulnérables

Lorsque le nouveau-né prématuré survit, il est considéré comme « vulnérable », c'est-à-dire nécessitant un suivi plus rapproché, particulièrement dans les premiers jours de vie. Le degré de vulnérabilité de l'enfant augmente avec la précocité de sa naissance. Sur le plan clinique, les enfants dits « vulnérables » sont les grands prématurés et les prématurés extrêmes (ou prématurissimes), ainsi que les nouveau-nés ayant un très faible poids de naissance (<1,500 g) ou présentant d'autres pathologies fœtales ou néonatales pouvant altérer leur développement futur. Les deux principaux groupes d'enfants vulnérables considérés dans cette thèse sont les grands prématurés et ceux à très faible poids de naissance.

Les naissances grandes prématurées ou de très faible poids sont les principales causes de morbidité et de mortalité périnatales dans le monde (4,10,11,42,43). Ces enfants vulnérables sont plus à risque de développer des troubles cognitifs et/ou moteurs à l'adolescence et à l'âge adulte (44–51). Le risque de développement des malformations et/ou des altérations cardiaques ou rénales est également accru (52). Des pathologies chroniques telles que l'asthme sont plus présentes chez ces enfants (53). Des troubles du comportement, de

l'attention, de l'apprentissage ou encore de la gestion des émotions sont également plus courants durant l'enfance chez les enfants vulnérables (54–57). Plusieurs études ont relevé que l'inquiétude et l'impuissance face au développement atypique de leurs enfants étaient plus importantes chez les parents d'enfants vulnérables que chez les parents d'autres enfants (58–61). Ces problèmes cliniques et comportementaux peuvent aussi avoir des conséquences pour les enfants qui auront d'autant plus de difficultés à s'intégrer socialement (à l'école, dans les activités périscolaires ou, plus tard, dans le monde professionnel). La naissance prématurée est également associée à des coûts importants pour les systèmes de santé (62–64). En plus d'un problème de santé publique, la question du développement des enfants vulnérables est également de fait un problème sociétal. La poursuite des études sur ce sujet est ainsi toujours nécessaire pour mieux comprendre l'ensemble des problématiques qui gravitent autour de la question de la prématurité.

1.4. Importance de la mise en place d'un suivi

1.4.1. Prise en charge des enfants vulnérables

Les anomalies de développement des enfants vulnérables peuvent être anticipées, améliorant ainsi la prise en charge de ces patients afin d'éviter des handicaps trop lourds à l'adolescence et à l'âge adulte (65–67). Cette anticipation permet aux professionnels de santé suivant l'enfant de mieux adapter les soins et la stratégie thérapeutique à adopter. Cela permet également de mieux anticiper les conséquences avec et sur les familles et de préparer au mieux avec elles la prise en charge de leur enfant (60,68). De plus, parmi tous les enfants vulnérables, tous ne seront pas concernés par des problèmes de développement. Mais seule une surveillance attentive par des professionnels médicaux formés et spécialistes de la question sur le long terme permet une telle constatation d'absence d'anomalie de développement.

Les recommandations internationales préconisent de fait un suivi clinique des enfants vulnérables jusqu'à l'âge de sept ans (69). Au cours de ce suivi les professionnels de santé se doivent de surveiller en particulier le développement moteur et cognitif de l'enfant. Un suivi rapproché lors de visites médicales annuelles voire pluriannuelles est par conséquent

nécessaire. Les professionnels médicaux et paramédicaux, spécialisés en pédiatrie, sont ainsi mobilisés pour assurer ce suivi. Au cours des visites, en plus des mesures anthropométriques de routine (poids, taille et périmètre crânien), ils analysent plusieurs dimensions du développement de l'enfant. Les examens sont adaptés en fonction de l'âge de l'enfant mais on y retrouve régulièrement des informations sur le développement de la motricité fine, de la cognition, de la vision, de l'audition, du langage, de l'expression orale et non verbale mais aussi sur d'autres items tels que le comportement, l'attention, l'alimentation, le sommeil ou encore l'appréciation des parents sur leur enfant. Ces examens se présentent le plus souvent sous la forme de questionnaires administrés par le professionnel de santé. Des auto-questionnaires standardisés, à compléter par les familles, sont également disponibles. Ils permettent à celles-ci de refléter le plus fidèlement possible l'état de santé perceptible au quotidien de l'enfant. Parmi eux, le questionnaire appelé *Age State Questionnaire* (ASQ) est reconnu internationalement pour évaluer le développement des enfants vulnérables à un âge donné (70). L'existence de telles pratiques permet non seulement aux médecins d'être le plus juste possible quant au diagnostic mais aussi garantit une comparaison valide entre tous les enfants participant à ce type de suivi. C'est ainsi que, dans les études épidémiologiques sur les enfants grands prématurés ou de très faible poids de naissance, ces variables standard sont très utiles pour établir l'état de santé de l'enfant.

L'état de santé clinique de l'enfant n'est pas le seul facteur à prendre en compte lorsqu'on étudie le suivi des enfants vulnérables. Des caractéristiques sociologiques, économiques, démographiques et géographiques s'ajoutent aux caractéristiques cliniques lors des études épidémiologiques des enfants vulnérables. Ainsi, à chaque mise en place de cohorte, cet aspect pluridimensionnel doit être pris en compte. Cependant, prendre en compte tous les paramètres importants pour le suivi des enfants vulnérables, et ce sur une durée suffisamment longue pour assurer un suivi complet, est parfois compliqué et coûteux. C'est pourquoi beaucoup d'études se concentrent sur une seule dimension (16,66,71–73) ou sur une courte période, généralement jusqu'à 2 ans (44,71,73).

1.4.2. Problématiques liées au suivi des enfants vulnérables

En parallèle des difficultés de mise en place, les programmes de suivi d'enfants vulnérables sont surtout lourds et fastidieux pour les enfants et les familles participants. Malgré tout, en plus des recommandations internationales qui préconisent de suivre ces enfants jusqu'à l'âge de sept ans, plusieurs études ont montré que la mise en place de ces programmes était efficace pour prévenir la survenue de problèmes de développement (69,71). L'anticipation et la surveillance permettent ainsi une prise en charge précoce des enfants afin d'améliorer leur état de santé actuel et à venir (65,66,69). Bien que, dans les cohortes de recherche (cf. tableau S1 en annexe 1), un effort soit effectué sur le suivi des patients pour limiter le nombre de perdus de vue, le taux d'attrition dans les cohortes d'enfants vulnérables est souvent important (74,75). Dans la cohorte française d'enfants vulnérables LIFT (76), le taux de perdus de vue était de plus de 10 % à 2 ans et de plus de 20 % à 5 ans. Dans une cohorte similaire, EIPAGE (77), le taux de perdus de vue à 5 ans était de 19 %.

Cet important taux d'attrition cause plusieurs problèmes. Tout d'abord les enfants perdus de vue sont potentiellement des enfants sans surveillance médicale et donc susceptibles de développer de lourdes anomalies à l'âge adulte. Ensuite, à l'échelle populationnelle, ce sont des manques d'informations multiples pour comprendre les problématiques des enfants vulnérables. Enfin il a été prouvé à plusieurs reprises qu'un taux d'attrition trop important peut conduire à des biais dans les études épidémiologiques (78). Ces différents constats justifient la nécessité de 1) fournir un effort humain, matériel et/ou organisationnel supplémentaire pour limiter au maximum le nombre de perdus de vue et 2) limiter au maximum les biais associés durant la phase d'analyse en utilisant une méthodologie statistique adaptée.

Concernant les cohortes d'enfants vulnérables, plusieurs études ont mis en exergue l'importance des taux d'attrition, leurs conséquences et les facteurs de risque associés. Parmi les facteurs de risque identifiés, on retrouve des facteurs de risque cliniques, socio-économiques mais aussi géographiques. Parmi les facteurs cliniques, les enfants nés avec une vulnérabilité moins importante, c'est-à-dire avec un moindre degré de prématurité, un plus grand poids de naissance ou de meilleures conditions néonatales, sont généralement plus à risque d'être perdus de vue (66,73–75). Même si l'on pourrait penser que ces enfants en

meilleure santé sont moins à risque de développer des anomalies durant leur jeune âge, ils n'en demeurent pas moins susceptibles de développer des anomalies tout au long de leur enfance. Leur absence de suivi médical est donc synonyme d'une perte d'anticipation sur la survenue d'anomalies futures.

Plus problématique, à statut clinique égal (y compris pour les enfants déjà très fragiles à la naissance), on retrouve des taux d'attrition plus élevés chez les enfants issus de familles précaires avec un faible niveau socio-économique. De nombreuses études de la littérature ont en effet montré qu'un bas niveau d'étude de la mère, de faibles revenus économiques, une instabilité professionnelle des parents, un logement trop petit pour la taille de la famille ou encore l'absence de couverture sociale étaient tous des facteurs de risque prédominants dans l'arrêt du suivi d'enfants vulnérables (71,73,77,79). Dans ces familles, la venue d'un enfant vulnérable dans le foyer est une contrainte supplémentaire qui vient s'ajouter à une multiplicité d'autres. Bien que la préoccupation du suivi de ces enfants soit présente dans ces familles précaires, le coût économique, la complication des procédures, la contrainte des rendez-vous, l'absence de communication autour de l'importance du suivi ou encore la barrière de langue sont autant de facteurs limitants dans le suivi. En plus de leur vulnérabilité clinique, les enfants issus de familles précaires souffrent ainsi d'une vulnérabilité sociale et leur défaut de suivi est un lourd préjudice dans leur développement. Cela peut entraver les efforts pour améliorer les soins et les interventions destinés à ces enfants.

De plus, indépendamment des facteurs socio-économiques, la situation géographique, c'est-à-dire le lieu de vie des familles, joue également un rôle prépondérant dans le suivi. Un éloignement trop important du lieu de suivi, des contraintes de déplacement trop lourdes (transports en commun) ou encore des particularités territoriales telles que des zones urbaines à très haute densité populationnelle ou au contraire des zones rurales à très faible densité favorisent la perte de suivi (66,79,80). Là encore persistent les occasions manquées de détecter précocement les anomalies de développement.

Pour finir, l'organisation du système de santé qui assure l'identification et le suivi des enfants vulnérables joue aussi un rôle primordial sur le nombre de perdus de vue. Opérer à une échelle populationnelle un tel suivi a un coût économique élevé et conditionne souvent la nature et le volume des moyens mis en place. De nombreuses caractéristiques associées à l'organisation de cette prise en charge ont ainsi souvent été identifiées comme des facteurs

associés au niveau d'attrition dans les populations d'enfants vulnérable (méthodologie de suivi, outils d'évaluation utilisés, nature et effectif des professionnels impliqués, particularités territoriales de la prise en charge, mise en place d'un système de planification et relances...) (71).

L'ensemble des paramètres, cliniques, socio-économiques, géographiques et organisationnels doivent ainsi être pris en compte lorsqu'on souhaite étudier et comprendre l'attrition dans les cohortes d'enfants vulnérables.

2. Médecine personnalisée et suivi des enfants vulnérables

Au début des années 2000, la notion de médecine personnalisée est apparue (81). Également connue sous le nom de *precision or personalized medicine* en anglais, elle a considérablement gagné en importance au cours de ces dernières années et les articles abordant cette thématique se sont multipliés (82–88).

La médecine personnalisée vise à améliorer la prévention et la gestion clinique des maladies en offrant aux patients des traitements spécifiquement adaptés et hautement ciblés (82). Pour atteindre cet objectif, cette démarche adopte une approche multidimensionnelle qui va au-delà de la simple considération de la pathologie. Elle intègre des données socio-environnementales, biopsychosociales, génétiques, biologiques, phénotypiques, ainsi que les caractéristiques spécifiques de la maladie (figure 3). L'ensemble de ces données forme une empreinte unique propre à chaque individu, créant ainsi une signature individuelle complexe qui caractérise l'organisme dans sa globalité.

Cette masse d'informations individuelles, spécifique à chaque patient, est ensuite exploitée par les professionnels de santé pour définir et adapter au mieux la prise en charge médicale ou les mesures de prévention à mettre en place pour chaque patient.

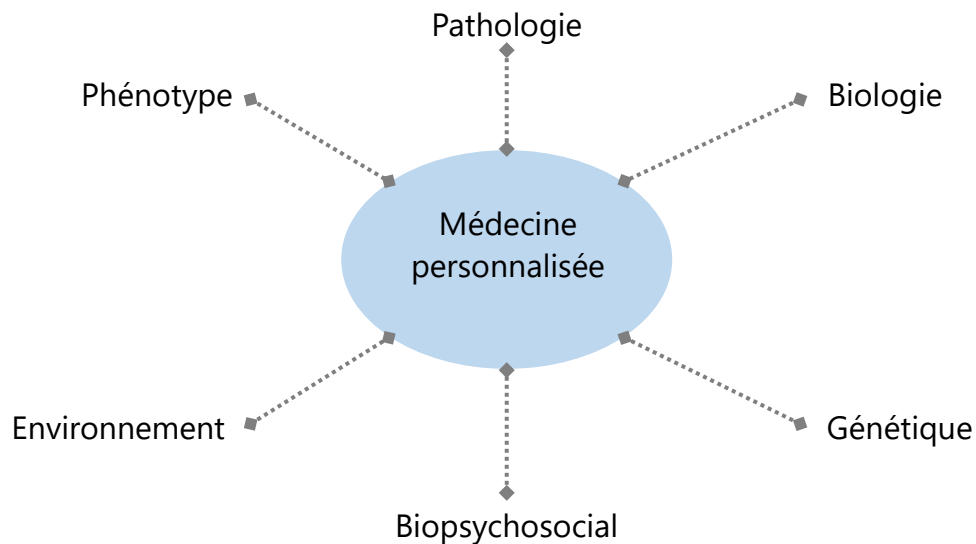


Figure 3 : Typologie des informations individuelles utilisées en médecine personnalisée

Au fil du temps, d'autres terminologies ont vu le jour pour affiner la philosophie et les objectifs de la médecine personnalisée : médecine 4P dans un premier temps (89–91), suivie par la médecine 5P (92,93). Avec l'émergence de la santé numérique et l'approche multidisciplinaire de la prise en charge médicale (94,95) on parle même de médecine 6P (81,95,96). La médecine 6P est une approche de la médecine personnalisée qui se caractérise par l'intégration de six principaux domaines d'analyse et d'intervention dans la prise en charge médicale et la gestion de la santé des individus. Chacun de ces domaines commence par la lettre P :

1. **Personnalisée** : La personnalisation des soins médicaux implique l'adaptation précise des traitements et des interventions médicales à chaque patient en fonction de ses caractéristiques individuelles. Cela peut inclure des facteurs tels que le profil génétique, l'historique médical, les préférences du patient et d'autres données spécifiques à chaque individu.
2. **Préventive** : La médecine préventive vise à réduire les risques de maladies et de complications en identifiant et en agissant sur les facteurs de risque potentiels. Elle se base sur des données de santé pour cibler des populations à risque élevé et mettre en place des mesures préventives adaptées.
3. **Prédictive** : La prédiction en médecine consiste à anticiper les risques de développer certaines maladies ou conditions de santé en se basant sur des informations

généétiques, des marqueurs de risque et d'autres données médicales. Cela permet une intervention précoce et ciblée pour prévenir ou atténuer ces risques.

4. **Participative** : La participation active du patient dans le processus de prise de décision médicale est encouragée. Cela signifie que les patients sont impliqués dans les choix liés à leur santé, y compris les options de traitement, les décisions sur les soins et la gestion de leur propre santé.
5. **Pertinente** : La pertinence en médecine se réfère à l'adoption de décisions de traitement et de soins qui tiennent compte des informations spécifiques à chaque patient, y compris son style de vie, son profil génétique, sa profession, et d'autres facteurs individuels. Ces décisions sont basées sur des preuves scientifiques solides.
6. **Parcours de soins** : La prise en charge médicale est multidisciplinaire, impliquant une équipe de professionnels de santé de diverses spécialités qui collaborent pour assurer une gestion globale et coordonnée de la santé du patient. Cette approche prend particulièrement en compte les aspects médicaux, sociaux, et professionnels de la vie du patient.

Cette réussite est intrinsèquement liée à la disponibilité accrue de données de santé nécessaires à ce type d'approche, conjuguée à l'utilisation récente de méthodologies de traitement avancées.

La collecte, l'accumulation et le traitement de données ont toujours été des activités humaines, bien avant l'avènement de la civilisation numérique (97). Cependant, les apports de données précédents semblent désormais modestes par rapport à ce qui est observé actuellement. En effet, le volume de données créées ces dernières années connaît une croissance exponentielle à l'échelle mondiale, passant de 2 zettaoctets (ou 2 milliards de téraoctets) en 2010 à 18 zettaoctets en 2016 et 64 zettaoctets en 2020. Ce chiffre pourrait atteindre, selon les prévisions, 181 zettaoctets en 2025 (98). Ce phénomène est souvent désigné par le terme de données massives, *big data* en anglais (99). La définition du *big data*, bien que très utilisée, n'est pourtant pas si évidente et plusieurs études se sont penchées sur la question (100–102). Bien qu'un socle commun existe sur l'emploi du terme lorsque les études statistiques se basent sur des données à forte volumétrie et de différents types (données structurées et non structurées), les conclusions sont unanimes sur le fait que la

définition de *big data* n'est pas bien borné. Elle repose fréquemment sur la notion des 3V (103) :

1. **Volume** : Le premier V fait référence à la dimension du volume des données. Dans le contexte du *big data*, les données sont générées en quantités importantes, bien au-delà de ce que nous avons connu par le passé. Ce volume massif de données peut provenir de diverses sources, notamment les réseaux sociaux, les capteurs, les systèmes de surveillance, les transactions en ligne etc
2. **Variété** : Le deuxième V concerne la diversité des données. Les données massives ne sont pas seulement volumineuses, elles sont également très diverses en termes de formats, de types et d'origines. Elles peuvent inclure des données structurées, telles que des bases de données, des données non structurées telles que des textes, des images et des vidéos, ainsi que des données semi-structurées, comme les données provenant de réseaux sociaux.
3. **Vélocité** : Le troisième V se rapporte à la vitesse des données, c'est-à-dire la rapidité à laquelle les données sont générées, collectées et traitées. Dans le monde du *big data*, les données sont souvent produites en temps réel ou à une fréquence très élevée. Par exemple, les réseaux sociaux génèrent en permanence des données, la transmission d'informations instantanées circulant à travers les internet, et les transactions financières sont effectuées en fractions de seconde

Pour éviter toute confusion et mésusage du terme *big data*, nous nous cantonnerons dans cette thèse à l'utilisation du terme « données massives » pour évoquer les données issues des outils numériques dans le contexte de la médecine personnalisée.

De nombreux domaines sont désormais touchés par cette abondance de données massives, allant de l'astronomie (104) à la finance (105,106), en passant par la publicité (107,108), la recherche scientifique (109), l'industrie manufacturière (110), la politique (111), et bien d'autres encore (112). La santé et la recherche médicale n'échappent pas à ce phénomène.

Les sources de données en santé se multiplient : données recueillies, numérisées lors d'une hospitalisation ou d'une consultation médicale (diagnostic, traitement, examen médical ou biologique, d'imagerie...), données de consommation de soins (données enregistrées par l'assurance maladie, prescriptions), données d'objets connectés – qu'ils soient médicaux

(tensiomètre, implant cardiaque) ou non médicaux à l'origine (pèse-personne, montre connectée, smartphone), et virtuellement toute autre source de données qui pourraient, dans un usage donné, être utiles : données socio-économiques, géographiques...

Le numérique en santé a ainsi permis d'accumuler une quantité de données de santé suffisamment importante et variée pour en exploiter le contenu et ainsi alimenter le processus de la médecine personnalisée (113) (1^{er} P de la médecine 6P). Grâce à l'utilisation des outils numériques qui viennent enrichir les informations sur son état de santé, le patient devient désormais acteur de sa propre santé, c'est ce qu'on appelle « l'expérience patient » (114) favorisant la médecine participative (4^{ème} P de la médecine 6P). Les données de santé du patient sont accessibles et partagées plus facilement par les professionnels de santé, ce qui leur permet de reconstituer et de bien coordonner le parcours de prise en charge des patients (6^{ème} P de la médecine 6P) et d'ainsi adapter au mieux la stratégie thérapeutique.

Dans une démarche préventive (2^{ème} P de la médecine personnalisée), le travail coopératif de l'ensemble de ces professionnels permet de renforcer le suivi individuel du patient. Le système de santé entier s'adapte aux besoins du patient avec toujours comme cible l'amélioration de son état de santé en prenant des décisions éclairées et pertinentes (5^{ème} P de la médecine 6P). En dernier lieu, l'un des objectifs essentiels de la médecine 6P réside dans l'exploration des données massives générées par les outils numériques, grâce à l'utilisation d'algorithmes statistiques à des fins prédictives (3^{ème} P de la médecine 6P). Cette démarche vise à mieux 1) comprendre, anticiper et prévenir l'état de santé des patients en tenant compte de toutes leurs caractéristiques et 2) définir, adapter et personnaliser le plus tôt possible un parcours de soins optimal pour chaque patient.

En pédiatrie et en santé périnatale, l'aspect prédictif de la médecine 6P a déjà fait l'objet d'une recherche approfondie (115–117), mettant principalement l'accent sur l'analyse des données omiques. Cependant, l'intégration récente du parcours de soins dans la médecine 6P a reçu moins d'attention jusqu'à présent. A notre connaissance, l'association de ces deux dimensions, à savoir le suivi du parcours de soins et la prédiction, au sein d'une même approche en pédiatrie et en santé périnatale demeure inexplorée. Cela confère à cette thèse une singularité et une nouveauté dans le domaine de la santé périnatale.

3. Algorithmes de prédiction

3.1. Définition générale et contexte d'utilisation

Une typologie des analyses statistiques en épidémiologie repose sur trois objectifs principaux : les analyses descriptives, explicatives, et prédictives.

Les analyses descriptives visent à décrire les caractéristiques des individus une à une ou en les croisant avec d'autres. Leurs objectifs en épidémiologie sont souvent de mesurer, dans une population sur une période donnée, la fréquence d'une pathologie (prévalence), le nombre de nouveaux cas de cette pathologie (incidence) ou le nombre de décès (mortalité), ou encore la prévalence d'une ou plusieurs expositions.

Les analyses explicatives font quant à elles appel à des modèles statistiques comme les régressions pour étudier l'association entre une variable à expliquer (Y) et une ou plusieurs variables explicatives (X). Elles sont très largement utilisées en épidémiologie car elles permettent de décrire ces associations (généralement force et direction de l'association) à l'aide de statistiques directement interprétables d'un point de vue épidémiologique. Ces dernières se basent en général sur l'utilisation de modèles s'appuyant sur des hypothèses (linéarité, homoscedasticité, additivité...) pour contraindre leur complexité et permettre l'estimation des paramètres d'intérêt.

La troisième approche est dite prédictive. Le but n'est plus d'étudier l'association entre X et Y mais de prédire la valeur de Y à l'aide des variables X . Pour réaliser ce type d'approche, une famille de méthodes s'appuyant sur des algorithmes de prédiction statistiques est de plus en plus utilisée. Le fonctionnement des algorithmes de prédiction repose sur l'utilisation de données précédemment observées pour prédire des événements non observés (118,119). Ces algorithmes sont par exemple utilisés en épidémiologie pour aider à la prise de décisions médicales ou de santé publique (119–122), la planification d'essais cliniques (123–125), la prédiction d'évènements ou du devenir pour les individus d'une population donnée (126,127).

3.2. Présentation des algorithmes de prédiction

Les algorithmes de prédiction sont nombreux et il existe plusieurs façons de les classer. Une méthode de classification consiste à les classer selon deux catégories : supervisés ou non supervisés (figure 4). Si le modèle est un algorithme supervisé, il peut être de deux types ou sous-catégories : algorithme de régression dans le cas où la variable de résultat Y est continue ou de classification dans le cas où Y est catégorielle.

Les algorithmes non supervisés sont utilisés pour tirer des conclusions et trouver des tendances à partir de données d'entrée. Cette catégorie de modèles désigne la situation où les données d'entrée X ne sont pas étiquetées par une variable Y (par exemple vivant ou mort). Il s'agit de découvrir les structures sous-jacentes à ces données. Les deux principales méthodes utilisées dans les algorithmes non supervisés comprennent le groupement (*cluster*) (128–130), la réduction de la dimensionnalité (131,132) ou encore les méthode de traitement automatique du langage ou « *Natural Language Processing* » (NLP) en anglais (133).

Les algorithmes supervisés consistent quant à eux à apprendre une fonction qui établit une correspondance entre des variables d'entrée X et de sortie Y . Par exemple, dans un ensemble de données avec deux variables, l'âge (entrée) et la taille (sortie), une méthode supervisée pourrait être mise en œuvre pour prédire la taille d'une personne en fonction de son âge.

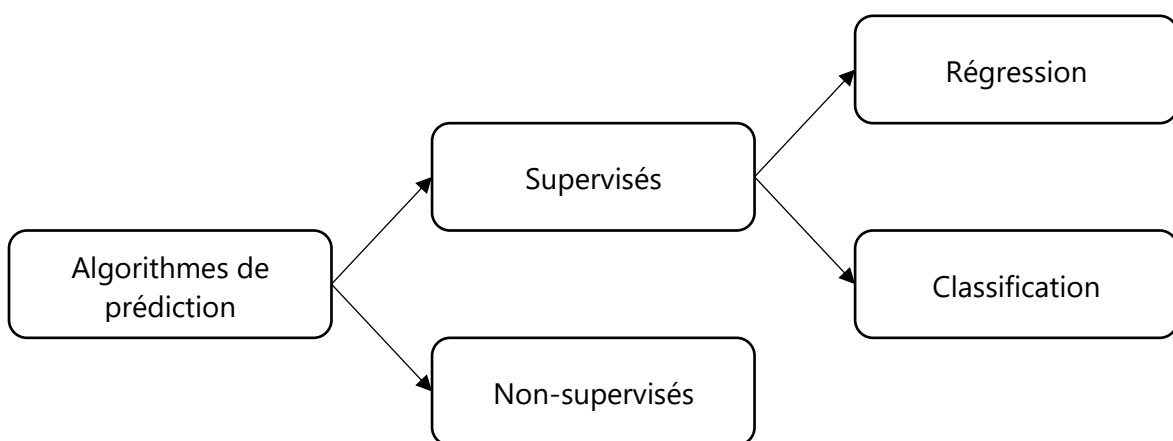


Figure 4 : Classification fondamentale des algorithmes de prédiction

Nous nous focaliserons dans les parties à venir sur la description des algorithmes de prédiction supervisés consacrés à la classification. L'objectif est alors de prédire la modalité de réponse (variable binaire Y) en fonction de variables prédictives X de toute nature (aussi bien catégorielles que continues). Chacun des algorithmes a des avantages et des inconvénients en fonction du contexte d'utilisation, des caractéristiques des données disponibles et des objectifs de l'analyse. Certains sont dits paramétriques, c'est-à-dire qu'ils supposent des hypothèses de départ comme la linéarité ou l'homoscédasticité, et d'autres sont dits non paramétriques, c'est-à-dire qu'ils ne supposent pas d'hypothèse de départ mais qu'ils doivent être définis par des paramètres fixés par le modélisateur. Ces paramètres sont généralement appelés « hyperparamètres » et influencent la qualité des modèles obtenus.

3.2.1. Régression logistique

La régression logistique (134) est un modèle supervisé classiquement utilisé en épidémiologie pour l'identification de facteurs associés (ou explicatifs) à une variable à expliquer. Concrètement elle modélise la relation supposée linéaire entre un vecteur de variables explicatives X et une fonction de la variable catégorielle à expliquer Y telle que la fonction logit (où $\text{logit}(Y) = \ln(P(Y=1)/P(Y=0))$). A l'aide de la méthode du maximum de vraisemblance (135), les coefficients de régression (souvent notés β) sont estimés pour identifier le modèle. On parle alors de modèle paramétrique. Dans le cas des analyses prédictives, l'utilisation du modèle de régression logistique, élaboré préalablement sur la base d'un échantillon de données, permet la prédiction de la variable Y à partir des données observées. Cette méthode a pour principal avantage d'être simple à mettre en place puisque, en estimant ses propres paramètres, elle ne dépend pas d'hyperparamètres à définir. De plus, la quantification de la relation entre les variables prédictives X et la variable à prédire Y permet une interprétation concrète des résultats. Cependant, comme toute méthode paramétrique, elle a pour principal inconvénient de dépendre d'hypothèses de base. En plus de supposer qu'il existe une relation linéaire entre les variables X et la variable Y transformée, on suppose également l'absence de colinéarité entre les variables X .

3.2.2. Régression LASSO

La régression *Least Absolute Shrinkage and Selection Operator* (LASSO) (136) est une méthode d'analyse de régression. Elle s'appuie sur de la sélection de variables et, en effectuant ce qu'on appelle une régularisation, elle limite le surajustement, c'est-à-dire le fait que le modèle s'adapte trop bien aux données d'entraînement au point de ne pas être en mesure de se généraliser à d'autres données du modèle final. Le but est d'améliorer les performances du modèle de régression logistique qu'elle produit. Elle permet, en partie, de limiter la multicolinéarité entre les variables prédictives, lorsque celles-ci sont nombreuses. En outre, elle permet également de réduire la complexité des modèles et donc la faisabilité de leur ajustement. La régression LASSO fonctionne de la même manière que la régression logistique classique mais introduit une contrainte supplémentaire reposant sur une technique de régularisation : la somme des valeurs absolues des coefficients de régression est pénalisée pour être aussi petite que possible. Le degré de cette pénalité est contrôlé par un paramètre de pénalité (appelé *penalty* en anglais) prédéfini. En augmentant la valeur de cette pénalité, certains coefficients de régression peuvent être réduits à zéro, ce qui a pour effet de sélectionner un sous-ensemble de variables prédictives (variables pour lesquelles le coefficient de régression associé reste différent de zéro). Ainsi, par rapport à la régression logistique classique, la régression LASSO fournit une sélection de variables permettant de mieux gérer la multicolinéarité, améliorant par ailleurs la prédiction.

Le modèle de régression LASSO est dit semi-paramétrique car il combine une méthode paramétrique (régression logistique) et utilise un hyperparamètre (***penalty***) à déterminer.

3.2.3. Régression multivariée par splines adaptatives (MARS)

La régression multivariée par splines adaptatives ou *Multivariate Adaptive Regression Splines* (MARS) a été proposée par Friedman en 1991 (137) pour modéliser des relations complexes entre les variables prédictives et la variable à prédire. Elle est utilisée pour analyser des données non linéaires. A l'aide de fonctions splines, fonctions polynomiales continues par morceaux, elle permet d'approximer les relations entre ces variables. Les points où ces fonctions changent sont appelés nœuds (ou *knots*). Par une méthode de sélection en avant

(*forward selection*), le modèle sélectionne un nombre de splines qui minimise une erreur de mesure. Puis, avec une procédure de sélection en arrière (*backward elimination*), il élimine les splines qui contribuent le moins à la précision du modèle. Cela aide à prévenir le surapprentissage et à produire un modèle qui se généralise mieux à de nouvelles données. Le degré des fonctions splines (quadratiques, cubiques...), appelé **degree**, ainsi que le nombre maximum de splines utilisées, appelé **num_terms**, sont les deux hyperparamètres à fixer pour cette approche.

3.2.4. Random Forests

En statistique, un arbre de classification est une méthode permettant de prédire le résultat d'une variable catégorielle à partir de la classification des observations effectuées sur une ou plusieurs variables prédictives. Le principe est détaillé dans le livre de Breiman et al. en 1984 (138). La figure 5 illustre le cas simple d'une variable à prédire Y binaire et de trois variables prédictives X_1 , X_2 et X_3 , toutes binaires. Le principe est que chaque nœud correspond au résultat d'une variable prédictive observée. Le résultat prédit sur la variable Y est obtenu lorsque la dichotomisation est suffisante pour l'arbre modélisé, c'est-à-dire lorsqu'il n'y a plus de nœud.

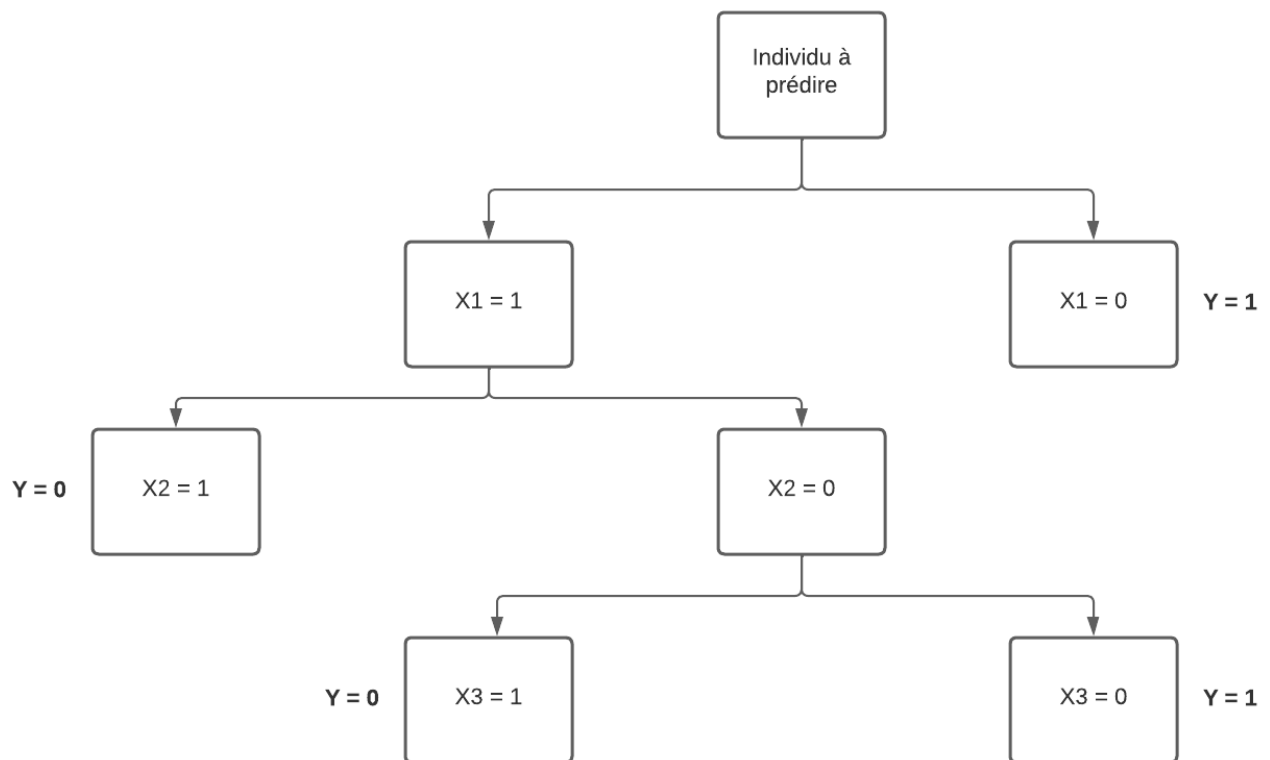


Figure 5 : Exemple d'un arbre de classification d'une variable à prédire binaire et de trois variables prédictives binaires

La méthode de *Bootstrap Aggregating (Bagging)* introduite par Breiman en 1996 (139) vise à réduire la variance des modèles de classification, en particulier pour les arbres de classification, en agrégeant les prédictions de plusieurs arbres formés de sous-ensembles de données légèrement différents. Pour créer ces sous-ensembles, la méthode effectue un tirage aléatoire avec remise des individus de l'échantillon de départ, obtenant ainsi plusieurs sous-échantillons de même taille que l'échantillon initial. Ainsi certains individus peuvent apparaître plusieurs fois dans un même sous-échantillon tandis que d'autres n'y figurent pas. Des arbres de classification sont ensuite entraînés indépendamment sur chacun de ces sous-échantillons. Plutôt que de faire des prédictions en utilisant un seul arbre, la méthode de *Bagging* prend en compte les prédictions de tous les arbres et utilise un vote majoritaire pour déterminer la prédiction finale. Comme décrit par Breiman, cette agrégation de prédictions de plusieurs modèles aide à réduire la variance et à améliorer la robustesse des prédictions par rapport à l'utilisation d'un arbre de décision unique.

Les algorithmes de forêts aléatoires ou *Random Forests* sont considérés comme une extension de la méthode de *Bagging* et ont été décrits par Breiman en 2001 (140). Pour résumer, le principe est de générer plusieurs arbres de classification en utilisant la méthode de *Bagging* puis d'ajouter une sélection aléatoire sur les variables prédictives à chaque nœud de l'arbre. Cela signifie qu'à chaque embranchement de l'arbre, au lieu d'évaluer toutes les variables prédictives pour déterminer la meilleure division, seul un sous-ensemble aléatoire de ces variables est considéré. De plus, pour éviter la trop grande complexité de l'arbre et ainsi prévenir le surajustement, un nombre minimum d'observations doit également être obtenu à chaque nœud.

On parle de « forêt d'arbres » en raison de la combinaison de plusieurs arbres induite par la méthode de *Bagging* et de « forêt aléatoire » en raison de l'utilisation de tirages aléatoires à plusieurs niveaux (sous-échantillon d'individus et sous-échantillon de variables). Par la suite nous utiliserons le terme anglais *Random Forests* majoritairement présent dans la littérature, plutôt que la traduction française.

Les *Random Forests* ne s'appuient pas sur des hypothèses de départ pour pouvoir être utilisées, elles sont donc considérées comme appartenant aux modèles non-paramétriques. Les trois hyperparamètres à définir sont le nombre de sous-échantillons tirés aléatoirement ou le nombre d'arbres dans la forêt, noté ***ntrees***, le nombre de variables sélectionnées aléatoirement à chaque nœud, ***mtry***, et le nombre d'observations minimales devant être présentes pour qu'un nœud soit divisé, aussi appelé profondeur de l'arbre ou ***min_n***.

3.2.5. Extreme Gradient Boosting (XGBoost)

Tout comme les *Random Forests*, l'eXtreme Gradient Boosting (XGBoost) (141) est une méthode basée sur un ensemble d'arbres de décision. Cependant, contrairement aux *Random Forests*, ces arbres ne sont pas générés indépendamment les uns des autres avec des sous-ensembles d'échantillons tirés aléatoirement. Dans la méthode de XGBoost, les arbres sont construits de manière séquentielle, en corrigeant les erreurs générées par l'arbre précédent. Pour ce faire, l'algorithme s'appuie sur une méthode d'amplification de gradient (*gradient boosting*) (142). Cette dernière méthode consiste à corriger les erreurs commises en sommant les prédictions de l'ensemble des arbres générés précédemment. De plus, les prédictions de

chaque nouvel arbre sont également pondérées par une pénalité appelée taux d'apprentissage qui est un hyperparamètre de l'approche noté **learn_rate**. Cette pénalité permet entre autres de réduire le surajustement du modèle. Pour optimiser le modèle, contrairement aux *Random Forests*, la profondeur de chaque arbre est limitée non plus seulement par l'absence de gain d'information significatif, mais aussi par un hyperparamètre fixé appelé **tree_depth**, qui correspond à la profondeur maximale de l'arbre. En d'autres termes, cela correspond au nombre de nœuds maximum présents dans chaque arbre.

Les hyperparamètres **ntrees** (nombre d'arbres), **mtry** (nombre de variables sélectionnées aléatoirement à chaque nœud) et **min_n** (nombre d'observations minimales devant être présentes pour qu'un nœud soit divisé) sont également les autres hyperparamètres qui restent à définir, tout comme pour les *Random Forests*.

3.2.6. Support Vector Machine (SVM)

Les machines à vecteurs de support ou séparateurs à vaste marge (*Support Vector Machine* ou SVM) sont une autre méthode d'apprentissage supervisé. Les SVM ont été développés dans les années 1990 par Cortes et Vapnik (143). Le but d'un SVM est de trouver l'hyperplan optimal qui sépare le mieux les différentes classes de données, c'est-à-dire celui qui maximise la marge entre les classes, la marge étant la distance entre l'hyperplan et les points de données les plus proches de chaque classe (figure 6). Dans cet exemple, les ronds verts correspondent au cas où $Y = 1$ et les carrés rouges au cas où $Y = 0$, X_1 et X_2 sont les deux variables prédictives continues. Dans cet exemple, des valeurs élevées à la fois pour X_1 et pour X_2 sont favorables au cas $Y = 1$, alors que des valeurs faibles de X_1 et X_2 sont davantage favorables au cas où $Y = 0$. L'hyperplan optimal est représenté par l'intérieur des deux diagonales dont la distance est égale à la marge du SVM.

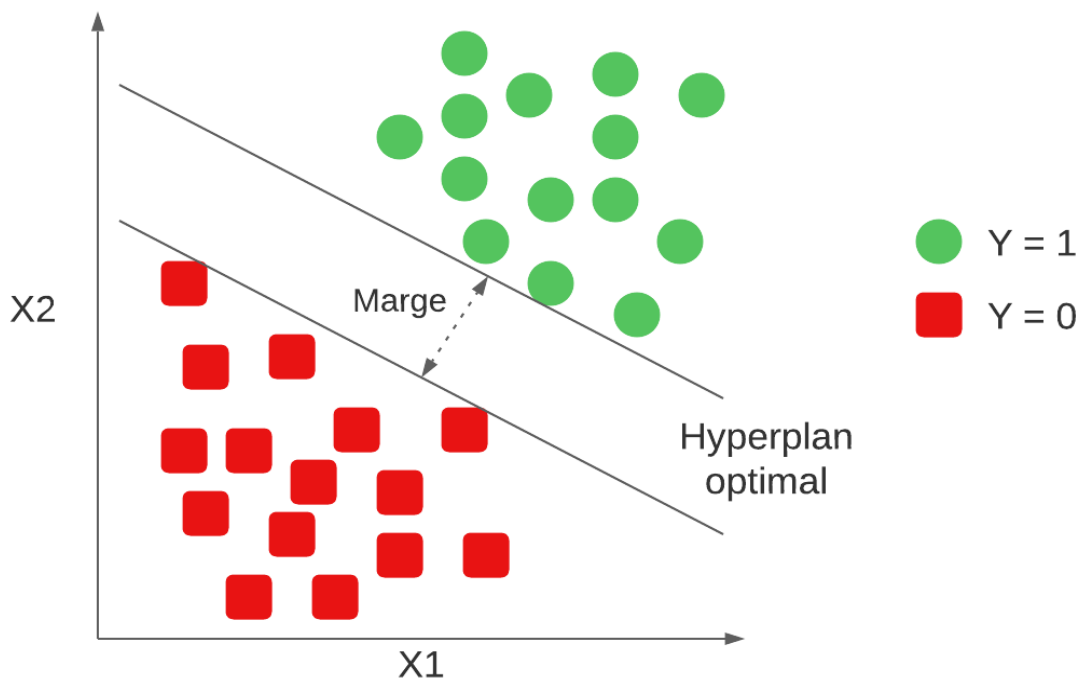


Figure 6 : Schéma simplifié de l'hyperplan linéaire d'un modèle de SVM

Dans les cas où les classes ne sont pas linéairement séparables, les données sont transformées dans un espace de plus grande dimension dans lequel elles peuvent être séparées linéairement. Ces fonctions sont appelées noyau ou *kernel* en anglais. Ces noyaux peuvent prendre diverses formes, comme les noyaux linéaires, polynomiaux et radiaux gaussiens ou *Radial Basis Function* (RBF). Le choix du noyau est important dans la paramétrisation d'un SVM puisqu'il permet d'identifier dans quel espace l'hyperplan optimal sera défini. Dans les trois cas, l'hyperparamètre **cost** est présent. Il s'agit du compromis entre les valeurs seuils autorisées des erreurs de classement et la maximisation de la marge. Concrètement c'est une pénalité associée aux erreurs de classement. Une pénalité élevée contribue à une plus grande importance aux erreurs de classement (et donc à une plus faible marge), ce qui aura pour conséquence d'obtenir un modèle plus discriminant, c'est-à-dire un modèle davantage en capacité de classer les données d'entraînement, mais avec un risque de surajustement aux données. Dans le cas d'un noyau RBF, le paramètre **rbf_sigma** correspond en quelque sorte à l'angle d'inclinaison du noyau Gaussien. Une valeur élevée donne un noyau large où chaque point d'échantillonnage a une influence plus faible mais s'étend sur une plus grande distance. Enfin dans le cas des noyaux polynomiaux, deux hyperparamètres sont à définir : le degré polynomial appelé **degree** qui correspond à l'espace de dimension du noyau, ainsi que le coefficient de facteur d'échelle, **scale_factor**, qui ajoute un coefficient sur les termes des polynômes. Un coefficient plus élevé correspond à une plus grande importance donnée aux termes de plus haut degré. Les hyperparamètres à définir sont donc dépendants du noyau choisi, comme résumé dans le tableau 3.

Tableau 3 : Hyperparamètres présents dans un SVM

Noyau	Nom	Définition
Linéaire	<i>cost</i>	Coût ou pénalité associée à chaque erreur de classification
Polynomial	<i>cost</i>	Coût ou pénalité associée à chaque erreur de classification
	<i>degree</i>	Degré du polynôme
	<i>scale_factor</i>	Coefficient de facteur d'échelle
Radial Basis Function	<i>cost</i>	Coût ou pénalité associée à chaque erreur de classification
	<i>rbf_sigma</i>	Largeur du noyau Gaussien

3.2.7. Réseaux de neurones

Les réseaux de neurones artificiels, ou plus simplement les réseaux de neurones, sont des modèles statistiques inspirés du fonctionnement des cerveaux biologiques (144,145). Dans un contexte de classification, ils utilisent des données d'entrée X (variables prédictives observées) pour prédire une variable de sortie Y (variable à prédire). Les données d'entrée sont introduites dans ce qu'on appelle une couche d'entrée, composée d'autant de neurones que de variables d'entrée du jeu de données. Chaque neurone représente ainsi un vecteur contenant les observations de l'échantillon. Chacun d'entre eux est ensuite connecté à des neurones d'une couche cachée. L'intensité de la connexion entre chaque neurone d'entrée et chaque neurone de la couche cachée est déterminée par des poids, initialement définis de manière aléatoire. En utilisant les sommes pondérées, la valeur de chaque neurone de la couche cachée est calculée. Ensuite, une fonction d'activation (logistique, tangente hyperbolique, ReLU, softmax...) est appliquée aux valeurs de chaque neurone de la couche cachée. De nouveaux poids sont attribués aux connexions entre les neurones de la couche cachée et ceux de la couche suivante ou de la couche de sortie si c'est la dernière couche cachée. Après avoir traversé toutes les couches cachées, des sommes pondérées sont utilisées pour obtenir la valeur de la variable à prédire Y dans la couche de sortie. L'objectif de l'entraînement est de minimiser une fonction de perte qui mesure la différence entre les prédictions du réseau et les valeurs réelles. Cela se fait en ajustant les poids des connexions à l'aide d'algorithmes d'optimisation tels que la descente de gradient stochastique (146). Le processus d'entraînement implique généralement de passer les données à travers le réseau plusieurs fois. Chaque passage est appelé **epoch**. Enfin, pour éviter le surapprentissage du modèle, une technique d'optimisation appelée **dropout** (147) est souvent utilisée. Durant chaque itération de l'entraînement, le **dropout** désactive de manière aléatoire certains neurones dans chaque couche (couche d'entrée et couches cachées). Cela signifie que ces neurones sont ignorés pendant cette itération spécifique de l'entraînement. Le pourcentage de neurones à désactiver dans chaque couche est un paramètre fixe à définir. Les neurones peuvent, d'un entraînement à l'autre, être inactif ou actif. Ainsi, cette technique force le réseau à apprendre des représentations plus robustes et indépendantes, puisqu'il ne peut pas se fier à la présence d'un neurone spécifique. Cela contribue à améliorer la capacité du modèle à généraliser à partir de nouvelles données non vues lors de l'entraînement.

Dans cet exemple, le réseau de neurones (figure 7) est constitué de trois variables prédictives (X_1, X_2, X_3) formant trois neurones dans la couche d'entrée. Chacun des trois neurones est connecté à cinq neurones (N_1, N_2, N_3, N_4 et N_5) présents dans la seule couche cachée du réseau. Chaque flèche représente le poids du lien entre les neurones de la couche d'entrée et les neurones de la couche de sortie. Si on note ω_1, ω_2 et ω_3 , les poids entre X_1, X_2, X_3 respectivement et le neurone N_1 , ce dernier est alors égal à $(X_1 * \omega_1) + (X_2 * \omega_2) + (X_3 * \omega_3)$. Il en est de même pour tous les autres neurones de la couche cachée. La fonction $f()$ représente la fonction d'activation appliquée à chaque valeur des neurones de la couche cachée, pour transformer la valeur des neurones. Sur le même principe, des sommes pondérées sont effectuées pour calculer la valeur finale de Y .

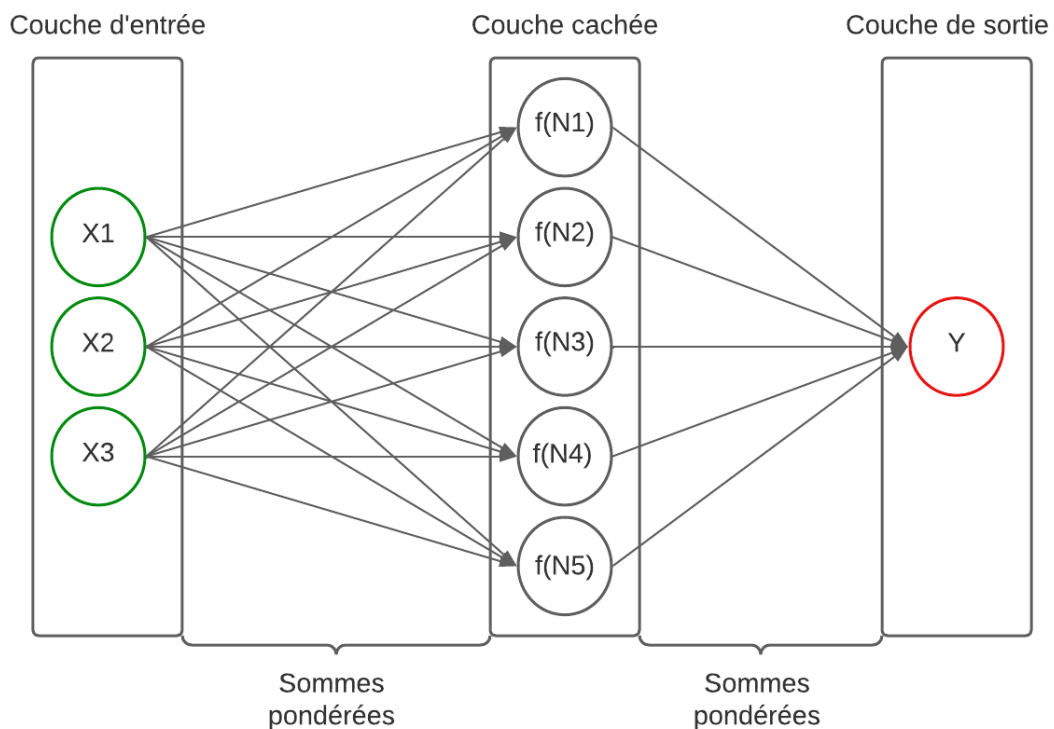


Figure 7 : Exemple d'un réseau de neurone à trois variables d'entrée, une couche cachée de cinq neurones et une variable de sortie binaire

Les réseaux de neurones sont complexes et leur paramétrisation doit être précise et adaptée au contexte des données. Les hyperparamètres à prendre en compte sont résumés dans le tableau 4.

Tableau 4 : Hyperparamètres présents dans un réseau de neurones

Nom	Définition
<i>hidden_layers</i>	Nombre de couches cachées
<i>hidden_units</i>	Nombre de neurones dans la couche cachée
<i>activation</i>	Fonction d'activation
<i>epochs</i>	Nombre d' <i>epochs</i> ou nombre de fois où le réseau est entraîné
<i>dropout</i>	Probabilité qu'un neurone soit inactif dans une couche à chaque entraînement

3.2.8. Méthode d'ensemble

Une méthode d'ensemble est une technique de modélisation statistique qui combine plusieurs algorithmes de prédiction pour améliorer les performances de prédiction. L'idée est que la combinaison de plusieurs modèles peut améliorer les performances globales en réduisant les erreurs individuelles de chaque modèle. Wolpert définit en 1992 la notion d'empilement ou *stacking* qui consiste à utiliser les résultats de plusieurs modèles de prédiction en entrée d'un nouveau modèle sur le jeu de données d'entraînement pour améliorer les performances des prédictions (148). La méthode *stack* a déjà prouvé son efficacité dans plusieurs études (123,149,150) car elle permet d'obtenir des résultats au moins aussi bons que chaque modèle pris à part. Cette méthode consiste à utiliser les prédictions des modèles un à un puis à les inclure comme entrée dans le modèle d'ensemble qui se base sur une régression linéaire pour déterminer des poids à chaque prédiction qui permettent de minimiser l'erreur de prédiction. En ajoutant une méthode LASSO à ce modèle d'ensemble, une pénalité (hyperparamètre **penalty**) sur chaque prédiction est également ajoutée pour ne sélectionner que les modèles de base qui ont les poids les plus importants dans les prédictions.

En revanche les méthodes d'ensemble peuvent être plus complexes et plus coûteuses que les modèles individuels, ce qui peut poser des défis pour la mise en œuvre en temps réel.

3.3. Optimisation d'un algorithme de prédiction

3.3.1. Pré-traitement des données

Dans un contexte de prédiction statistique, le pré-traitement de données (*data pre-processing* en anglais) joue un rôle crucial pour garantir la qualité et la fiabilité des prévisions (151). En effet, les données utilisées pour la prédiction sont susceptibles de contenir des anomalies liées à des valeurs aberrantes, des doublons, des données manquantes, des incohérences, etc. Le pré-traitement de données consiste à nettoyer, à transformer et à préparer ces données. Cette étape inclut la correction des erreurs, l'imputation des données manquantes, la normalisation et la standardisation des variables, la sélection des variables, la réduction de la dimensionnalité, etc. En outre, l'objectif principal du pré-traitement de données est d'améliorer la qualité des données pour maximiser la précision et la performance des modèles de prédiction utilisés car il garantit que les données d'entrée soient propres, homogènes et bien préparées pour l'analyse et la prédiction.

Les principales étapes du pré-traitement des données sont les suivantes

1. La correction des erreurs et des valeurs aberrantes constitue une étape cruciale lors du prétraitement des données, en particulier dans le cas d'un jeu de données de vie réelle. Souvent, ces données comportent des erreurs de saisie ou des valeurs aberrantes qui peuvent influencer de manière significative les résultats des prédictions statistiques. Pour garantir la qualité des données en entrée, il est essentiel d'adopter une approche rigoureuse pour corriger ces valeurs aberrantes. L'une des méthodes courantes pour repérer les valeurs aberrantes dans les variables quantitatives est de définir des bornes cohérentes en fonction du domaine d'étude. Par exemple, lorsqu'on considère l'âge de patients enfants, il peut être pertinent de définir une plage normale entre 0 et 15 ans. Si une donnée dépasse ces bornes, plusieurs approches sont envisageables en fonction du contexte. On peut choisir de considérer la donnée comme manquante si l'écart est minime, ou bien la modifier pour qu'elle corresponde à la borne la plus proche (par exemple, rapporter l'âge d'un patient de 16 ans à 15 ans). Une autre option pourrait être de supprimer directement l'observation si l'écart est trop important pour être justifié. Il est important de souligner que les décisions relatives à la correction des valeurs aberrantes doivent être prises en concertation avec les cliniciens, les

investigateurs de l'étude et en tenant compte des connaissances existantes dans la littérature. En effet, chaque cas est unique, et les choix opérés doivent être en adéquation avec le contexte spécifique de l'étude. Une attention particulière portée à cette étape garantit des prédictions plus robustes et mieux adaptées au contexte de l'étude, tout en évitant les biais potentiellement induits par des données erronées ou aberrantes.

2. L'exploitation et la transformation de variables préexistantes dans le jeu de données initial constituent également une étape du pré-traitement de données. Ces méthodes permettent d'extraire des informations plus spécifiques à partir d'une seule variable et de créer de nouvelles variables, augmentant ainsi la puissance et la précision du modèle de prédiction. Par exemple, à partir d'une variable de date, on peut isoler l'année, le mois ou le jour pour examiner l'effet calendaire ou saisonnier. De même, on peut catégoriser une variable continue, c'est-à-dire la diviser en classes distinctes pour la rendre plus interprétable. Ces transformations facilitent la compréhension des modèles et contribuent à révéler des tendances et des schémas sous-jacents qui peuvent être cruciaux pour la prédiction statistique.

3. L'imputation des données manquantes est une étape majeure dans le processus de mise en place d'un algorithme de prédiction statistique. En effet, la plupart des algorithmes ne supportent pas les données manquantes et suppriment automatiquement les observations concernées. Comme pour les données aberrantes, le nombre de données manquantes dans un jeu de données réel peut être important. Il est donc crucial de les prendre en considération afin de ne pas perdre d'information. L'imputation de données consiste à remplacer les valeurs manquantes par des valeurs probablement proches de la réalité. Plusieurs méthodes d'imputation existent, telles que l'imputation multiple et l'imputation simple. L'imputation multiple (152), souvent évitée dans le cadre de la prédiction statistique en raison de sa complexité, consiste à générer plusieurs jeux de données imputés et à appliquer l'algorithme sur chaque jeu. L'imputation simple, plus simple et moins coûteuse en termes de temps de calcul, consiste à remplacer la valeur manquante par une valeur supposée être proche de la réalité. Plusieurs méthodes d'imputation simple existent, telles que l'imputation par le mode (pour les variables qualitatives), par la moyenne ou la médiane (pour les variables quantitatives), ou encore par la technique de plus proches voisins, ou *k-Nearest Neighbors* (*k*-NN) (153). Cette dernière méthode calcule la valeur imputée en fonction d'un nombre

déterminé de "voisins" (paramètre k), qui sont des individus ayant des valeurs observées proches de celles de l'individu possédant des valeurs manquantes. Pour éviter de complexifier le processus d'imputation ou pour éviter de générer des biais potentiels dus au trop grand nombre de données manquantes, il est également concevable d'exclure avant imputation les variables ayant un trop grand nombre de données manquantes. Une autre possibilité est de considérer la donnée manquante comme une information en tant que telle et donc d'imputer ces données manquantes en modalité de type « inconnu » et évaluer les capacités prédictives a posteriori.

4. Suppression des variables peu ou pas informatives. Une variable sans variation, c'est-à-dire ayant une variance égale à zéro (avec la même valeur pour toutes les observations), n'apporte aucune information prédictive et doit donc être supprimée du jeu de données. De même, pour réduire la complexité du modèle, les variables dont la variance est proche de zéro peuvent être écartées, car elles contribuent peu à la prédiction. En outre, les variables fortement corrélées entre elles peuvent fournir une information redondante. Dans ce cas, une seule d'entre elles peut être conservée, tandis que les autres sont supprimées. Cette démarche contribue à alléger le modèle et à limiter le risque de surapprentissage, en éliminant les données qui n'améliorent pas la capacité prédictive du modèle.

5. Transformation des variables quantitatives. Certains algorithmes de prédiction font l'hypothèse que les variables prédictives suivent une distribution normale ou une relation linéaire avec la variable Y . Par conséquent, il peut être nécessaire de transformer les variables prédictives pour lisser les données afin qu'elles soient adaptées à l'algorithme choisi. Ces transformations peuvent inclure le centrage et la réduction des données (pour les ramener à une moyenne de zéro et une variance égale à 1), l'application d'une transformation de type Box-Cox (154), pour stabiliser la variance, ou l'utilisation de fonctions splines pour lisser et linéariser les données. Ces techniques contribuent à assurer que les données répondent aux exigences de l'algorithme, améliorant ainsi la qualité et la robustesse de la prédiction.

6. Création d'indicatrices. La plupart des algorithmes de prédiction ne prennent en compte que des variables numériques, il est donc important de transformer au préalable les variables qualitatives en indicatrices. Cela implique la création d'une variable binaire pour chaque modalité, qui prend la valeur 0 si l'observation ne présente pas la modalité, et 1 si elle

la présente. Cette technique de transformation est connue sous le nom de « *one hot encoding* ». Elle permet de convertir les attributs catégoriels en une forme que l'algorithme peut manipuler, tout en conservant l'information essentielle contenue dans la variable originale.

7. Gestion des données déséquilibrées. Lorsque, dans le jeu de données initial (données d'entraînement), le déséquilibre est trop important entre les deux classes de Y (dans le cas de la classification binaire), il existe des méthodes pour corriger ce déséquilibre de classe. De manière générale ces méthodes consistent à créer des observations synthétiques de la classe minoritaire de Y en générant des valeurs aléatoires pour les variables prédictives X .

3.3.2. Optimisation des hyperparamètres

Une grille de recherche est une technique d'optimisation d'hyperparamètres qui consiste à énumérer systématiquement un ensemble de valeurs pour chaque hyperparamètre d'un modèle de prédiction statistique et à évaluer les performances du modèle pour chaque combinaison de valeurs. Son utilisation contribue grandement à l'amélioration des algorithmes de prédiction (155).

Considérons par exemple un modèle de prédiction statistique à trois hyperparamètres (A , B et C) et que nous choisissons de tester trois valeurs pour A , deux valeurs pour B et quatre valeurs pour C . La grille de recherche testera alors l'ensemble des $3 \times 2 \times 4 = 24$ combinaisons possibles de valeurs pour ces hyperparamètres sur le jeu d'entraînement. Pour chaque combinaison de valeurs, le modèle est entraîné et évalué sur un ensemble de données de validation. Les performances du modèle sont ensuite enregistrées pour chaque combinaison d'hyperparamètres, et la combinaison avec les meilleures performances est sélectionnée pour le modèle final. En généralisant, si l'on considère N combinaisons, chacune d'entre elles sera testée et évaluée sur les données d'entraînement. Pour chacune d'entre elles une métrique à déterminer sera donc donnée. En admettant que la métrique de la combinaison i donne la meilleure valeur, ce sera donc cette combinaison d'hyperparamètres qui sera choisie et

implémentée dans le modèle final. La figure 8 présente un schéma explicatif du fonctionnement de la grille de recherche.

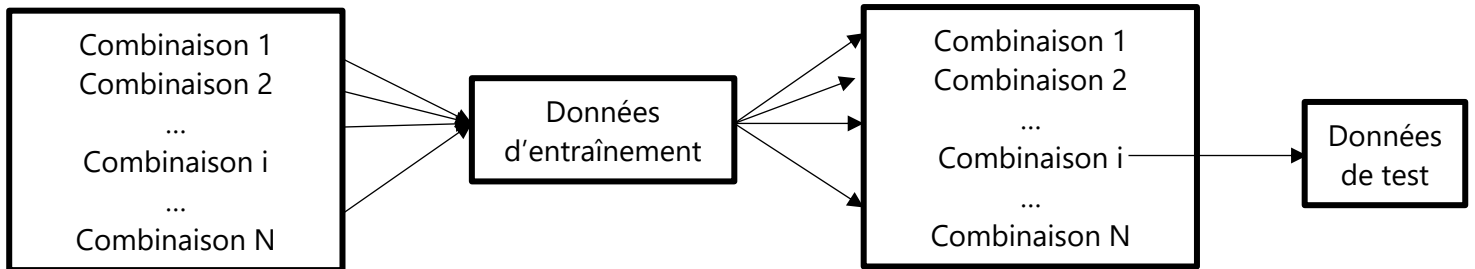


Figure 8 : Utilisation des combinaisons pour la grille de recherche

La grille de recherche est une méthode simple et efficace pour trouver les meilleures combinaisons d'hyperparamètres pour un modèle de prédiction statistique. L'inconvénient est qu'elle peut être très coûteuse en temps de calcul si le nombre d'hyperparamètres et le nombre de valeurs à tester pour chaque hyperparamètre sont élevés. C'est pourquoi il est important de trouver un juste milieu entre optimisation et temps de calcul.

Il existe plusieurs types de grilles de recherche :

1. Grilles aléatoires : les grilles aléatoires sélectionnent les valeurs des paramètres de manière aléatoire à partir d'une distribution prédéfinie. Cette méthode permet d'explorer un grand espace de paramètres de manière efficace mais elle peut être moins systématique que d'autres méthodes de recherche de paramètres car elle ne permet pas d'explorer l'ensemble de l'espace défini.
2. Maximum d'entropie : cette méthode utilise la théorie de l'information pour sélectionner les valeurs des paramètres qui maximisent l'entropie du modèle. L'idée est que le modèle le plus général est celui qui a le plus d'entropie, et donc le plus de complexité. Cette méthode peut être efficace pour les modèles très complexes mais elle peut être coûteuse en temps de calcul.
3. Hypercube : cette méthode divise l'espace des paramètres en une grille régulière et teste toutes les combinaisons de paramètres dans chaque cellule de la grille. Cette méthode est efficace pour les modèles avec un petit nombre

de paramètres mais elle peut devenir impraticable pour les modèles avec de nombreux paramètres.

4. Optimisation bayésienne : cette méthode utilise des techniques d'optimisation bayésienne pour sélectionner les valeurs des paramètres qui maximisent la probabilité du modèle, en utilisant une fonction de coût définie. Cette méthode est particulièrement efficace pour les modèles qui ont des paramètres complexes ou non linéaires.

3.3.3. Validation croisée

La validation croisée (*cross-validation*) est une technique d'évaluation de la performance d'un modèle d'apprentissage automatique en utilisant l'ensemble des données disponibles (156). Elle consiste à diviser les données en K sous-ensembles d'individus (ou *folds*), en général de taille égale, et à répéter le processus d'entraînement et de validation du modèle K fois. À chaque itération, $K-1$ sous-ensembles sont utilisés pour entraîner le modèle, tandis que le sous-ensemble restant est utilisé pour le tester. Cette technique permet d'éviter le surapprentissage (ou *overfitting*) du modèle sur les données d'entraînement et de mieux évaluer la performance du modèle sur les données de test et les nouvelles données.

La validation croisée stratifiée permet de résoudre les problèmes de classification où les classes sont déséquilibrées. Elle s'assure que chaque sous-ensemble contient un nombre similaire d'instances de chaque classe.

La validation croisée est une technique importante pour évaluer la performance des modèles d'apprentissage automatique, notamment dans le choix des hyperparamètres du modèle et pour estimer la performance attendue sur de nouvelles données.

3.4. Evaluation de la performance

Pour évaluer la performance d'un modèle, la définition au préalable de métriques à maximiser est nécessaire. Selon le contexte, les objectifs et le type de variables à prédire, le choix de la ou des métriques à maximiser (ou minimiser selon les cas) et à évaluer peut varier.

Pour évaluer la performance globale d'un algorithme de classification, on s'appuie sur la matrice de confusion. Une matrice de confusion est un tableau comparant les prédictions de l'algorithme avec les valeurs observées. Elle permet de visualiser les résultats de la classification en montrant le nombre de prédictions correctes et incorrectes pour chaque classe. La matrice de confusion est généralement de taille $k \times k$, où k est le nombre de classes. Les colonnes de la matrice représentent les valeurs réelles et les lignes représentent les prédictions de l'algorithme. Les éléments diagonaux de la matrice correspondent aux prédictions correctes pour chaque classe, tandis que les éléments hors diagonale correspondent aux prédictions incorrectes.

A partir de cette matrice de confusion, différentes métriques peuvent être calculées. Prenons le cas simple d'un nombre de classes égal à deux avec comme variable à prédire la survenue d'une maladie (Positif = cas malade, Négatif = non malade). Nous obtenons la matrice de confusion (figure 9) avec en colonne les classes observées du jeu de données et en ligne les classes telles que prédites par l'algorithme. VP correspond ainsi aux Vrais Positifs (individus réellement malades et détectés comme tels), FN correspond aux Faux Négatifs (individus réellement malades mais détectés comme non-malades), FP correspond aux Faux Positifs (individus non malades mais détectés comme malades) et enfin VN correspond aux Vrais négatifs (individus non malades et détectés comme tels).

		Observations	
		Positif	Négatif
Prédiction	Positif	VP	FP
	Négatif	FN	VN

Figure 9 : Exemple d'une matrice de confusion à deux entrées

A partir de cette matrice de confusion on obtient entre autres les métriques suivantes :

- Sensibilité (ou rappel) = $VP/(VP+FN)$
- Spécificité = $VN/(FP+VN)$
- Valeur Prédicative Positive (VPP) (ou précision) = $VP/(VP+FP)$
- Valeur Prédicative Négative (VPN) = $VN/(FN+VN)$

Ainsi la sensibilité mesure la capacité de l'algorithme à détecter effectivement les cas malades. La spécificité mesure quant à elle la capacité de l'algorithme à détecter effectivement les non-malades. La VPP mesure la capacité de l'algorithme à identifier les cas malades et enfin la VPN mesure la fiabilité de l'algorithme sur les individus détectés comme non-malades.

Pour prendre en compte l'ensemble de ces paramètres, une autre métrique (*Accuracy*) permet d'évaluer la précision globale de l'algorithme sans mesurer spécifiquement la sensibilité et la spécificité mais en ayant l'avantage de considérer les deux dans une même mesure. Elle se calcule avec la formule suivante :

$$Accuracy = (VP+VN)/(VP+FN+FP+VN).$$

Une autre métrique pour les cas d'algorithme de classification binaire est également régulièrement utilisée. Elle est dérivée de la courbe ROC (pour *Receiver Operating Characteristic*) (157) qui représente la relation entre le taux de VP (Sensibilité) et le taux de FP (1 - Spécificité) pour différents seuils de classification. La courbe ROC est obtenue en traçant le taux de vrais positifs en fonction du taux de faux positifs pour différents seuils de classification. Plus le modèle est performant, plus la courbe ROC sera proche du coin supérieur gauche du graphique. Une courbe ROC idéale correspond à la ligne verticale reliant les points (0,0) et (0,1) complétée par la ligne horizontale reliant les points (0,1) et (1,1). Ce qui représente un taux de vrais positifs de 1 et un taux de faux positifs de 0 pour tous les seuils de classification. Une courbe ROC médiocre représente la diagonale reliant les points (0,0) et (1,1). Elle correspond à une répartition de probabilité égale entre les positifs et les négatifs et donc une répartition complètement aléatoire. Un exemple de courbe ROC est représenté sur la figure 10.

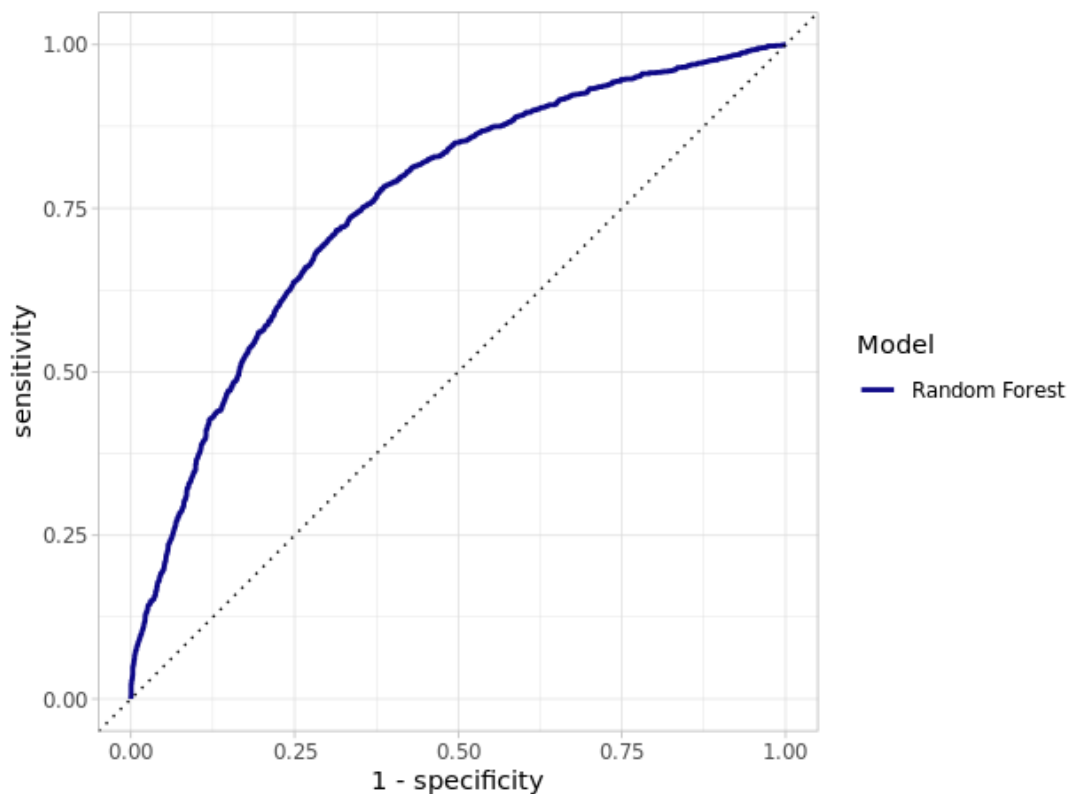


Figure 10 : Exemple d'une courbe ROC

L'aire sous la courbe ROC ou *Area Under the Curve* (AUC) est la métrique à maximiser pour chaque modèle à utiliser puisqu'elle permet directement d'évaluer les capacités du modèle à classer les positifs et les négatifs. Dans l'exemple ci-dessus la valeur de l'AUC est de 0,74. D'autres types de courbes que la courbe ROC existent comme la *precision-recall* qui représente la sensibilité en fonction de la VPP.

A noter qu'il existe d'autres métriques calculables à partir de la matrice de confusion comme le score $F1 = 2 * (VPP * Sensibilité) / (VPP + Sensibilité)$ mais nous nous focaliserons uniquement sur les métriques présentées ci-dessus dans le reste de ce manuscrit.

Il est important de choisir une ou plusieurs métriques appropriées pour évaluer la performance du modèle en fonction de l'objectif du projet et des caractéristiques des données.

Ce ne sera pas détaillé ici mais, tout comme pour les algorithmes de classification, il existe plusieurs métriques d'évaluation couramment utilisées pour les algorithmes de prédiction d'une variable continue (algorithmes de régression).

4. Objectifs de recherche

Le présent travail de thèse vise à développer des outils afin d'améliorer la prise en charge des enfants grands prématurés. Pour y répondre trois objectifs spécifiques ont été identifiés : 1) La compréhension des problématiques liées au suivi des enfants vulnérables, 2) L'utilisation de techniques statistiques innovantes en santé et 3) Le renforcement du suivi individualisé.

4.1. Comprendre les problématiques liées au suivi des enfants vulnérables

En s'appuyant sur une cohorte d'enfants vulnérables, ce travail de thèse vise à décrire et identifier les facteurs associés à la perte de vue, permettant ainsi de mieux comprendre les mécanismes et les enjeux liés aux problématiques en lien avec le suivi des enfants vulnérables.

4.2. Utiliser des techniques innovantes pour améliorer la prise en charge des enfants vulnérables

Le deuxième objectif de cette thèse est de proposer des techniques innovantes en statistique pour améliorer la prise de décision des professionnels en charge du suivi des enfants vulnérables. Cette thèse vise donc également à développer des algorithmes de prédiction statistique et à en évaluer les performances pour construire des scores individualisés sur le devenir des enfants vulnérables.

4.3. Renforcer le suivi individualisé

En appliquant les résultats des modèles établis dans le cadre de la prédiction statistique aux enfants nouvellement inclus, le troisième objectif spécifique de cette thèse est de renforcer

le suivi individualisé. Ce troisième objectif entre dans le cadre de l'amélioration du parcours du patient à travers la médecine personnalisée.

Afin de répondre à l'ensemble de ces objectifs, ce mémoire se concentrera dans un premier temps sur la présentation de la cohorte HYGIE-SEV qui a été utilisée pour ce travail de thèse, en mettant en avant les aspects méthodologiques de sa mise en place (chapitre 2). Ensuite, seront abordés les aspects épidémiologiques permettant de déterminer les facteurs associés à la perte de vue chez les patients de cette cohorte (chapitre 3). La méthodologie employée et les résultats des algorithmes de prédiction sur le devenir des enfants vulnérables seront illustrés dans un troisième temps (chapitres 4 et 5). Enfin, dans le cadre d'une discussion générale, il sera convenu de souligner l'importance de l'approche adoptée ici dans le contexte de la médecine personnalisée (chapitre 6).

Par le biais de l'Association Nationale de la Recherche et de la Technologie (ANRT), cette thèse s'inscrit dans le cadre d'une Convention Industrielle de Formation par la Recherche (CIFRe) pour laquelle plusieurs enjeux industriels étaient envisagés. Le premier enjeu visait à accroître la visibilité des données provenant du programme HYGIE-SEV pour les investigateurs de l'étude et les médecins participants. Le second enjeu cherchait à explorer des voies de compréhension liées à la perte de vue, ce qui permettrait d'identifier de manière plus précise les populations à risque pour les personnes responsables du suivi. Enfin la construction de scores de prédiction, leur restitution et leur mise en production auprès des acteurs concernés par le programme HYGIE-SEV était également un des objectifs industriels. Afin de répondre efficacement à ces objectifs, les efforts déployés au cours de la réalisation de cette thèse se sont concentrés sur 1) l'assurance de la reproductibilité des résultats et leur utilisation en routine afin d'intégrer les nouvelles inclusions au fil de l'eau, 2) la communication auprès des investigateurs sous forme d'envoi de rapports automatisés, de présentations orales ou d'interface web, 3) l'optimisation de l'utilisation des ressources internes disponibles, telles que les serveurs de calcul, les entrepôts de données et les serveurs web et 4) une veille technologique active et pertinente en lien avec les travaux entrepris dans le cadre de cette thèse.

CHAPITRE 2 –

DONNEES : LE DISPOSITIF HYGIE-SEV

Ce deuxième chapitre présente le dispositif HYGIE-SEV à l'origine des données sur lesquelles s'appuient les travaux réalisés dans la thèse. Le protocole suivi de ce dispositif et les données collectées sont ainsi décrits dans ce chapitre. Ce chapitre présente également les résultats d'un premier travail visant à décrire les caractéristiques des enfants à leur inclusion dans le programme HYGIE-SEV.

1. Présentation de HYGIE-SEV

1.1. Contexte et justification du programme

Comme vu précédemment les enfants nés très prématurément ou avec un poids de naissance très faible courent un risque élevé de développer des troubles pendant l'enfance qui peuvent perdurer pendant l'adolescence et à l'âge adulte (45–51). Nous avons également vu que les recommandations internationales préconisent de suivre ces enfants dès leur sortie de l'hôpital jusqu'à l'âge de sept ans (69). Cependant ces suivis doivent s'adapter aux problématiques territoriales dans lesquels naissent et grandissent ces enfants. En France, chaque région administrative doit avoir un programme de suivi qui lui est propre (158). Le programme est ensuite piloté par un ou plusieurs Réseau(x) de Santé en Périnatalité (RSP) présent(s) dans la région (159). Un réseau de santé en périnatalité est une structure de coordination, d'appui, d'évaluation et d'expertise médicale exerçant des missions dans le champ de la santé périnatale (159). Dans l'actualisation de la circulaire de l'harmonisation des missions des réseaux de santé en périnatalité dans un cadre régional du 3 juillet 2015, il est également précisé que le réseau doit assurer ces missions au-delà de la période périnatale pour le Suivi des Enfants Vulnérables (SEV), au cours des premiers mois ou années de vie selon les cas, jusqu'à l'âge de six ans (159). A ce titre, le réseau associe également les

professionnels intervenant dans le champ de la santé de l'enfant au sens large et du handicap. La Fédération Française des Réseaux de Périnatalité (FFRSP) promulgue et accompagne chaque RSP (160). En communiquant de près avec les réseaux, en étant représentée de manière équitable par des membres de chaque réseau et en étant proche du terrain, elle maintient une harmonisation sur l'ensemble du territoire français tout en s'adaptant aux problématiques locales.

Dans la région Ile-de-France, il existe 6 RSP :

1. RPSOF : Réseau Pédiatrique Sud et Ouest Francilien
2. RSPP : Réseau de Santé Périnatal Parisien
3. NEF : Naître dans l'Est Francilien
4. RPVM : Réseau Périnatal du Val de Marne
5. RPVO : Réseau Périnatal du Val d'Oise
6. MYPA : Maternité en Yvelines et Périnatalité Active

Ces réseaux et leur localisation sont représentés sur la carte de la figure 11.

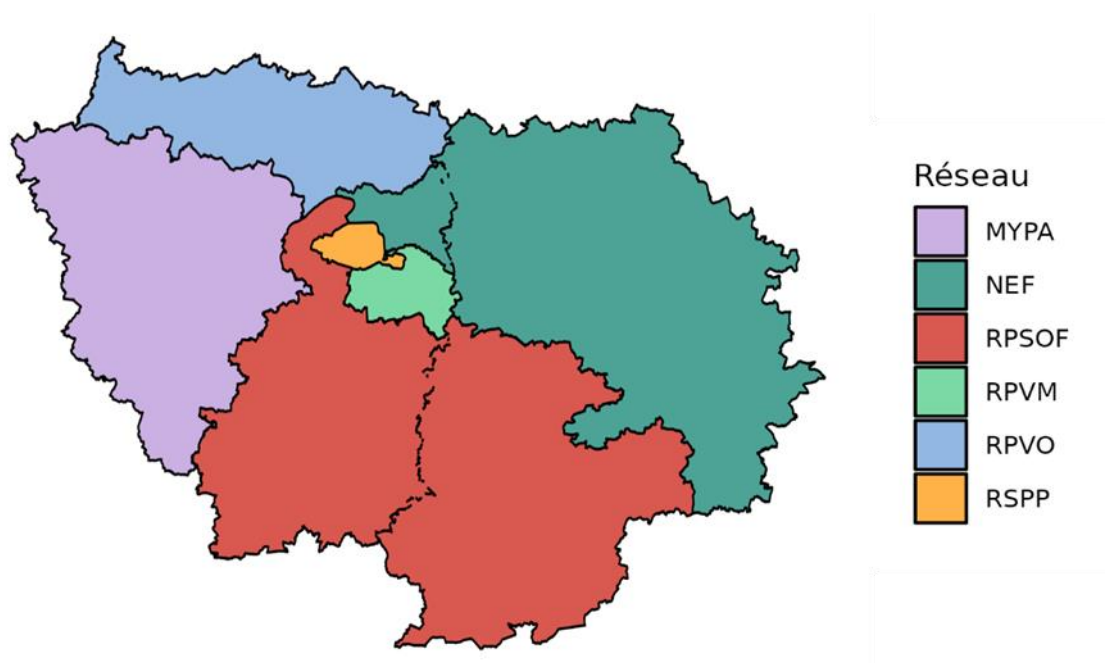


Figure 11 : Répartition des réseaux de santé en périnatalité en Île-de-France

L'ensemble de ces RSP constitue une offre de santé publique dont la coordination régionale, le pilotage et les financements sont assurés par l'Agence Régionale de Santé d'Ile-de-France (ARS-IDF). Les ARS ont été mises en place en 2010 dans une volonté de régionalisation des politiques de santé (161). Elles ont notamment pour rôle de réguler, financer et piloter l'offre de soins de chaque région administrative qu'elles gouvernent. Dans le contexte de la périnatalité et de la pédiatrie, chaque ARS s'occupe de financer et de coordonner le ou les RSP de sa région. En Ile-de-France, il s'agit donc de coordonner et de financer les 6 RSP. En incluant les 6 RSP, le réseau global du Suivi des Enfants Vulnérables est assuré par un réseau de 750 médecins pilotes et 78 maternités d'inclusion. Ces médecins pilotes regroupent un ensemble de professionnels de santé, avec entre autres des médecins spécialistes, des médecins généralistes, des professionnels paramédicaux (sages-femmes, kinésithérapeutes, psychomotriciens...), des nutritionnistes et des psychologues.

Pour assurer le suivi des enfants vulnérables, la conception de programmes d'intervention spécifiques efficaces (162) et le développement d'outils numériques et de méthodes innovantes ont déjà prouvé leur efficacité pour mieux aider les professionnels de santé dans ces tâches (163). L'ARS-IDF a ainsi décidé de mettre en place une solution numérique appelée HYGIE-SEV en collaboration avec les équipes de SESAN (164). SESAN est une entreprise privée à but non lucratif, partenaire privilégié de l'ARS-IDF sur toutes les questions de construction d'outils numériques et de systèmes d'information en lien avec l'amélioration de la coordination et de la coopération sanitaire. L'interface web HYGIE-SEV est un dossier patient partagé par les professionnels de santé qui interviennent dans le suivi des enfants vulnérables de moins de 7 ans. Elle permet notamment d'inclure de nouveaux patients, de compléter des formulaires de visite pour chaque patient, de consulter et de modifier les fiches. Les données collectées comprennent entre autres des informations cliniques, sociologiques et économiques. Le programme HYGIE-SEV a été testé en phase pilote pendant près de 10 ans dans certains RSP d'Ile-de-France et a été officiellement lancé le 1^{er} novembre 2015. Le recrutement des enfants vulnérables dans le programme HYGIE-SEV est toujours en cours à ce jour. Avec l'appui de son département de traitement de l'information, SESAN a également en charge les analyses relatives aux données de HYGIE-SEV. Un comité scientifique relatif à HYGIE-SEV a également été mis en place en 2020. Ce comité a pour principal objectif d'améliorer la qualité du suivi des enfants inclus dans le programme HYGIE-SEV. Il est

composé de différents corps professionnels : investigateurs du programme (ARS-IDF), néonatalogistes, coordinateurs des RSP, chercheurs associés et *data scientists* de SESAN en charge du traitement de données de HYGIE-SEV.

1.2. Objectifs du programme

Le programme HYGIE-SEV vise à atteindre les objectifs globaux suivants : détecter le plus précocement possible les anomalies du développement chez les enfants jusqu'à l'âge de sept ans, comprendre leurs mécanismes et identifier leurs facteurs de risque, fournir des données épidémiologiques pour informer et guider les professionnels de santé dans leur démarche de prévention, évaluer et optimiser le suivi de routine des enfants à haut risque dans la région, ainsi que rassembler un groupe multidisciplinaire d'experts afin de coordonner et développer la recherche épidémiologique et opérationnelle dans le domaine de la périnatalité.

2. Protocole de suivi des enfants vulnérables

D'un point de vue épidémiologique, les données produites par le programme HYGIE-SEV peuvent être considérées comme une cohorte prospective multicentrique ouverte basée sur des données que l'on qualifiera dans la suite de cette thèse de données de vie réelle. Bien que ce dispositif HYGIE-SEV peut être considéré comme une intervention, l'exploitation des données issues d'un logiciel « métier » utilisé dans les pratiques quotidiennes des professionnels donne un caractère « vie réelle » de la cohorte qui en découle. Les enfants inclus sont suivis de la naissance jusqu'à l'âge de sept ans.

2.1. Critères d'inclusion du programme

Les enfants éligibles à l'inclusion dans le programme HYGIE-SEV sont ceux qui sont en vie à leur sortie de l'une des 78 maternités de la région et dont l'âge gestationnel est inférieur à 33 semaines ou dont le poids de naissance est inférieur à 1 500 grammes. Parmi ces maternités, 15 disposent d'une unité néonatale avec soins intensifs (figure 12).

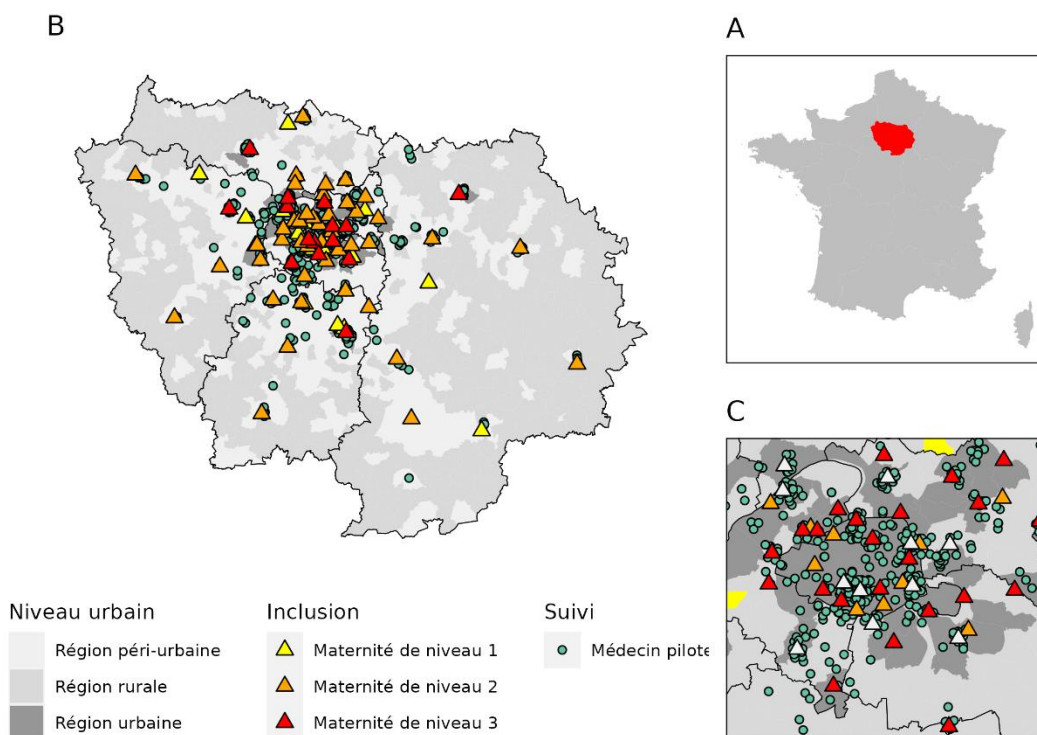


Figure 12 : Répartition des lieux d'inclusion (maternités) et de suivi (médecins pilotes) en perspective du niveau urbain dans HYGIE-SEV, région Île-de-France, 2015-2022

2.2. Suivi des enfants inclus

A sa sortie d'hospitalisation, chaque enfant est inclus dans un établissement (qui peut être différent de son établissement de naissance) et un médecin pilote lui est attribué. Ce médecin pilote a pour charge de réaliser les visites de suivi annuelles et de saisir les informations sur le logiciel HYGIE-SEV.

Après l'inclusion, l'enfant est revu à quatre mois d'âge corrigé. L'âge corrigé correspond à l'âge de l'enfant s'il était né à terme. Dans HYGIE-SEV, le terme étant fixé à 37 SA soit 257 jours, l'âge corrigé se calcule selon la formule suivante :

$$\text{Age corrigé (en jours)} = \text{Age réel de l'enfant} - (257 - \text{Terme de l'enfant})$$

L'enfant est également revu à un an et deux ans d'âge corrigé. A partir de trois ans, l'enfant est revu autour de chacun de ses anniversaires (âge réel) jusqu'à l'âge de sept ans. Ainsi, dès l'inclusion, chaque date de visite est d'ores et déjà programmée automatiquement dans le

logiciel afin de faciliter le suivi. Sous la forme d'une frise chronologique (*timeline*), le médecin peut ainsi avoir une vision globale des visites programmées. La figure 13 illustre un exemple de visuel pour le médecin pilote d'un enfant sélectionné.

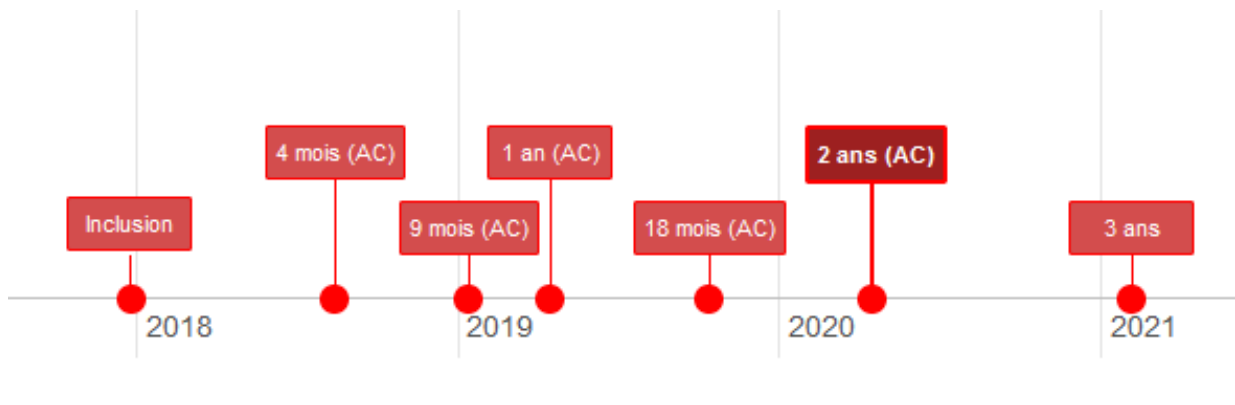


Figure 13 : Exemple d'une timeline dans HYGIE-SEV

En cliquant sur une visite, un formulaire associé et propre à chaque visite s'ouvre et le médecin peut ainsi remplir toutes les informations correspondant aux caractéristiques de l'enfant. Dans les parties ci-dessous, sont détaillées de manière plus approfondie les différentes données collectées.

2.3. Plateforme informatique

L'outil numérique permettant l'utilisation du programme de HYGIE-SEV se présente sous forme d'une plateforme informatique. A travers cette plateforme, les données recueillies, constituent un système d'information complexe de parcours patient en périnatalité. La récupération des données « brutes » fut la première étape du travail de thèse. Au total, ce sont 462 tables dans la seule base de données de HYGIE-SEV. Ces tables de données ne contiennent pas toutes des informations utiles pour les questions abordées ici. On retrouve des données non structurées (PDF, images...) qui servent à construire l'outil et la structure de l'outil informatique ou encore des données « système » utiles pour l'industriel ayant réalisé cette solution numérique. En se servant du manuel d'exploitation de l'architecture technique à notre disposition (cf annexe 2), une première sélection de tables de données nécessaires à

l'exploitation des données de HYGIE-SEV a été réalisée en premier lieu. Ainsi un panel de 16 tables de données a été utilisé pour construire toutes les analyses réalisées dans ce travail de thèse. Les informations issues de ces tables concernaient des informations sur les patients inclus, sur les formulaires de suivi, des informations sur les professionnels de santé participants ou encore des informations sur l'organisation du réseau SEV. Ainsi les tables de suivi collectées variaient de la simple table de correspondance, par exemple une table de données de 25 lignes et deux colonnes pour la table de correspondance des professions des professionnels de santé, à une table de 19 millions de lignes à ce jour et cinq colonnes pour les informations saisies dans chaque formulaire de suivi. Le traitement de ces données massives fut un véritable premier défi pour faciliter et fluidifier les analyses permettant de répondre aux questions de recherche qui ont été posées ici. Tout en veillant à bien récupérer les bonnes informations au bon endroit, il a également fallu être attentif aux erreurs inhérentes au SI. Par exemple, en cas de doublons de visite, la visite la plus récente était conservée. Pour structurer les données en tables d'analyse, nous avons fait le choix de conserver une table par visite obligatoire (4 mois, 1 an, 2 ans, 3 ans, 4 ans, 5 ans, 6 ans et 7 ans) regroupant les informations des patients recueillies durant la visite si celle-ci avait été réalisée, une table comprenant les informations du formulaire des données générales (informations socio-économiques des foyers, date de naissance de la mère, conditions de logement,...), une table concernant l'adresse rattachée à l'enfant (plusieurs adresses sont possibles si celui-ci a déménagé), une table contenant les informations des médecins pilotes, une table contenant les données recueillies à l'inclusion et une table renseignant les sorties de réseaux. Au total ce sont 13 tables de données qui ont été créées et qui ont servi de tables d'analyse pour tous les travaux présentés ici.

La création des tables d'analyse seule ne suffit pas à exploiter directement les données. Les étapes de *data management* sont essentielles dans toute étude épidémiologique en particulier dans un contexte de *data science* comme c'était le cas ici. Ces étapes sont là pour assurer un nettoyage de données nécessaire pour :

1. Une organisation pertinente du schéma de données afin de faciliter et fluidifier le flux de données
2. Corriger des problèmes de cohérence dans les données. Par exemple, sur HYGIE-SEV, certaines saisies de poids étaient en grammes et d'autres en kilogrammes, il a donc fallu convertir le tout dans une même unité de mesure.
3. Enrichir les données avec d'autres sources, par exemple ici les données de l'INSEE sur le nombre d'habitants par commune (165)
4. Calculer des scores composites à partir de variables existantes ou encore sélectionner les individus et les variables qui seront utiles pour les analyses statistiques. Les Z-scores de Fenton ont par exemple été calculés pour les données de poids, taille et périmètre crânien à la naissance et à la sortie (166).

3. Données collectées

Un grand nombre de données sont collectées à la naissance, à la sortie de l'hôpital et au cours du suivi pour tous les enfants inclus. Le tableau 5 présente un résumé des données collectées dans le cadre du programme HYGIE-SEV.

Tableau 5 : Données collectées dans le programme HYGIE-SEV

Données collectées	Age (m = mois, a = années)								
	Inclusion	4 m*	1 a*	2 a*	3 a	4 a	5 a	6 a	7 a
Caractéristiques socio-économiques									
Date de naissance (enfant)	X								
Lieu de résidence (enfant)	X	X	X	X	X	X	X	X	X
Structure familiale (enfant)	X	X	X	X	X	X	X	X	X
Date de naissance (mère)	X								
Statut marital (mère)	X	X	X	X	X	X	X	X	X
Couverture sociale (mère)	X	X	X	X	X	X	X	X	X
Conditions de logement (mère)	X	X	X	X	X	X	X	X	X
Niveau d'éducation (parents)	X	X	X	X	X	X	X	X	X
Catégorie socio-professionnelles (parents)	X	X	X	X	X	X	X	X	X
Source de revenus (parents)	X	X	X	X	X	X	X	X	X

Tableau 5 (suite)

Données collectées	Age (m = mois, a = années)								
	Inclusion	4 m*	1 a*	2 a*	3 a	4 a	5 a	6 a	7 a
Caractéristiques de la grossesse et néonatales									
Mode d'accouchement	X								
Accouchement provoqué	X								
Type de naissance (simple ou multiple)	X								
Procréation Médicalement Assistée	X								
Complication pendant la grossesse	X								
Consommation de tabac/drogue/alcool pendant la grossesse	X								
Sexe de l'enfant	X								
Age gestationnel	X								
Données anthropométrique (poids, taille, périmètre crânien)	X								
Mode de garde prévu pour l'enfant à la sortie	X								
Allaitement	X								
Caractéristiques de la prise en charge médicale									
Maternité de naissance et d'inclusion	X								
Niveau de la maternité d'inclusion	X								
Transfert de maternité à la naissance	X								
Réseau en charge du suivi	X	X	X	X	X	X	X	X	X
Date de sortie d'hospitalisation	X								
Dates des visites de suivi réalisées		X	X	X	X	X	X	X	X
Caractéristiques des médecins pilotes	X								
Date et motifs de sortie du réseau	X	X	X	X	X	X	X	X	X
Evaluation neuromotrice	X	X	X	X	X	X	X	X	X
Evaluation sensorielle	X	X	X	X	X	X	X	X	X
Evaluation respiratoire	X	X	X	X	X	X	X	X	X
Données anthropométrique (poids, taille, périmètre crânien)	X	X	X	X	X	X	X	X	X
Interactions sociales		X	X	X	X	X	X	X	X
Evaluation du sommeil			X	X	X	X	X	X	X
Evaluation de la nutrition			X	X	X	X	X	X	X
Evaluation du comportement			X	X	X	X	X	X	X
Evaluation de la qualité de vie			X	X	X	X	X	X	X
<i>Age and Stage Questionnaire (ASQ)</i>				X					
Propreté diurne et nocturne				X	X	X	X	X	X
Evaluation des capacités d'attention					X	X	X	X	X
Scolarité						X	X	X	X
Mode de garde de l'enfant	X	X	X	X	X	X	X	X	X

* Age corrigé

3.1. Caractéristiques socio-économiques

Les caractéristiques socio-économiques sont évaluées à la sortie de l'hôpital et mises à jour à chaque visite de suivi à l'aide de questionnaires standardisés administrés par les médecins. Les données démographiques pour l'enfant comprennent la date de naissance, le lieu de résidence, la structure familiale, les conditions de logement, la couverture médicale et les ressources économiques du ménage. Les données démographiques de la mère comprennent l'âge de la mère, le statut marital, le niveau d'éducation et la catégorie socioprofessionnelle (selon la nomenclature de l'Institut National de la Statistique et des Etudes Economiques (INSEE) de 2003 (167)). La catégorie socioprofessionnelle du père est également recueillie.

3.2. Caractéristiques de la grossesse et du nouveau-né

Les données concernant la grossesse, telles que le mode d'accouchement, l'accouchement provoqué, la naissance multiple, l'assistance médicale à la procréation, les complications, la prescription anténatale de corticostéroïdes, la chorioamniotite, le tabagisme et la consommation de drogues pendant la grossesse, sont systématiquement collectées dans le programme HYGIE-SEV à la sortie de la maternité d'inclusion.

Les données recueillies systématiquement pendant la période néonatale comprennent le sexe, l'âge gestationnel, les paramètres anthropométriques (poids, taille et périmètre crânien) de l'enfant à la naissance et à la sortie de l'hôpital, ainsi que des informations sur l'état de l'enfant, les soins, la prise en charge et le traitement pendant l'hospitalisation initiale, tels que l'intervention chirurgicale à la naissance, la corticothérapie postnatale, la prescription d'anticorps monoclonaux contre le virus respiratoire syncytial (VRS), la dysplasie broncho-pulmonaire (nourrissons dépendants de l'oxygène à 36 SA), les anomalies neurologiques évaluées en neuro-imagerie (risque faible, modéré ou élevé détecté lors de l'évaluation neurologique), la rétinopathie de stade II ou supérieur et d'autres anomalies sensorielles ou pathologies fœtales.

3.3. Caractéristiques de la prise en charge médicale

Plusieurs données concernant la prise en charge médicale sont enregistrées, notamment : la maternité de naissance et de sortie s'il y a eu un transfert pendant ou avant l'hospitalisation, la situation géographique et le niveau de la maternité, le RSP responsable de l'enfant à la sortie, les caractéristiques du médecin assurant le suivi (y compris son statut hospitalier ou libéral, sa spécialité médicale et son lieu d'exercice), ainsi que toutes les dates clés du suivi (hospitalisation, sortie, visites de suivi, et fin du suivi). Les raisons de la fin du suivi sont également enregistrées lorsqu'elles sont disponibles.

Le médecin pilote procède à plusieurs évaluations cliniques du développement de l'enfant au cours de visites spécifiques, en se concentrant sur les paramètres anthropométriques (poids, taille et périmètre crânien) et en détectant toute anomalie respiratoire, sensorielle, de croissance et neuromotrice. Le médecin utilise également différents scores standardisés en fonction de l'âge pour évaluer les interactions sociales de l'enfant, ses comportements, sa capacité d'attention, la qualité de son sommeil et de son alimentation, ainsi que l'acquisition de la propreté et la qualité de vie en général. Des questions concernant la scolarité et les caractéristiques de la garde d'enfants sont également posées. Le contenu de chacun des scores développés spécifiquement pour le programme HYGIE SEV est détaillé en annexe 3. Ces scores se basent tous sur le même principe : il s'agit d'une échelle de 1 à 4 (parfois 5, parfois 3), la valeur 1 représentant le cas où l'enfant ne présente aucune anomalie de développement particulière pour le score en question et les valeurs les plus élevées correspondant aux anomalies les plus sévères pour le score considéré.

À l'âge de 24 mois, en plus des scores standardisés renseignés par le médecin pilote, les parents sont invités à remplir l'auto-questionnaire *Age and Stage Questionnaire* (ASQ) (168). L'ASQ est un questionnaire composé de questions spécifiques à l'âge, réparties en cinq domaines de capacités de l'enfant : capacités de communication, motricité globale, motricité fine, capacités de résolution de problèmes et aptitudes personnelles et sociales.

4. Etude descriptive : population et méthodes

Une analyse descriptive des caractéristiques des enfants à leur inclusion dans le programme HYGIE-SEV est présentée ici. Dans cette étude sont inclus tous les enfants participant au programme HYGIE-SEV depuis son lancement jusqu'en date de la dernière extraction de la base de données, soit le 31 juillet 2023. Pour plus d'homogénéité, nous avons choisi de nous focaliser sur les enfants grands ou extrêmes prématurés (naissance strictement avant 32 SA) plutôt que sur le critère d'inclusion de HYGIE-SEV qui concernait les enfants nés strictement avant 33 SA. Tous les enfants avec un poids de naissance de moins de 1 500 g (deuxième critère d'inclusion dans HYGIE-SEV) ont également été considérés dans cette étude descriptive. Les caractéristiques des autres enfants inclus ne sont pas décrites séparément car il s'agit d'une population hétérogène et dont les effectifs sont relativement limités.

Les données ont été décrites à l'aide d'effectifs et pourcentages pour les variables catégorielles, moyennes et écarts-types pour les variables continues, en comparant 3 populations : les enfants nés grands ou extrêmes prématurés, les enfants dont le poids de naissance est inférieur à 1 500 g et la somme de ces deux populations. A noter qu'un enfant pouvait à la fois être grand ou extrême prématuré et avoir un poids inférieur à 1 500 g. Pour cette raison, il n'a pas été effectué de test statistique pour comparer les deux populations qui n'étaient donc pas indépendantes.

5. Résultats descriptifs : caractéristiques à l'inclusion

Entre le 1^{er} novembre 2015 et le 31 juillet 2023, 17 296 enfants ont été inclus dans la cohorte HYGIE-SEV. Parmi ceux-ci, 14 557 ont été inclus dans l'étude de description des données, soit plus de 2 200 enfants chaque année, dont 12 334 (84,7 %) grands ou extrêmes prématurés (naissances avant 32 SA) et 12 007 (82,5 %) naissances de nouveau nés ayant un très faible poids de naissance (<1 500 g) (figure 14). Autrement dit, sur les 14 557 enfants inclus, 2 223 (15,3 %) n'étaient pas grands ou extrêmes prématurés mais avaient un poids de moins de 1 500 g contre 2 550 (17,5 %) étant grands ou extrêmes prématurés mais avec un poids de naissance supérieur à 1 500 g. Ces deux populations ne s'excluent pas mais le travail présenté

ici permet en outre de décrire chacun des deux groupes définis par un critère bien précis en plus de la description du groupe « total », lui-même étant un critère composite.

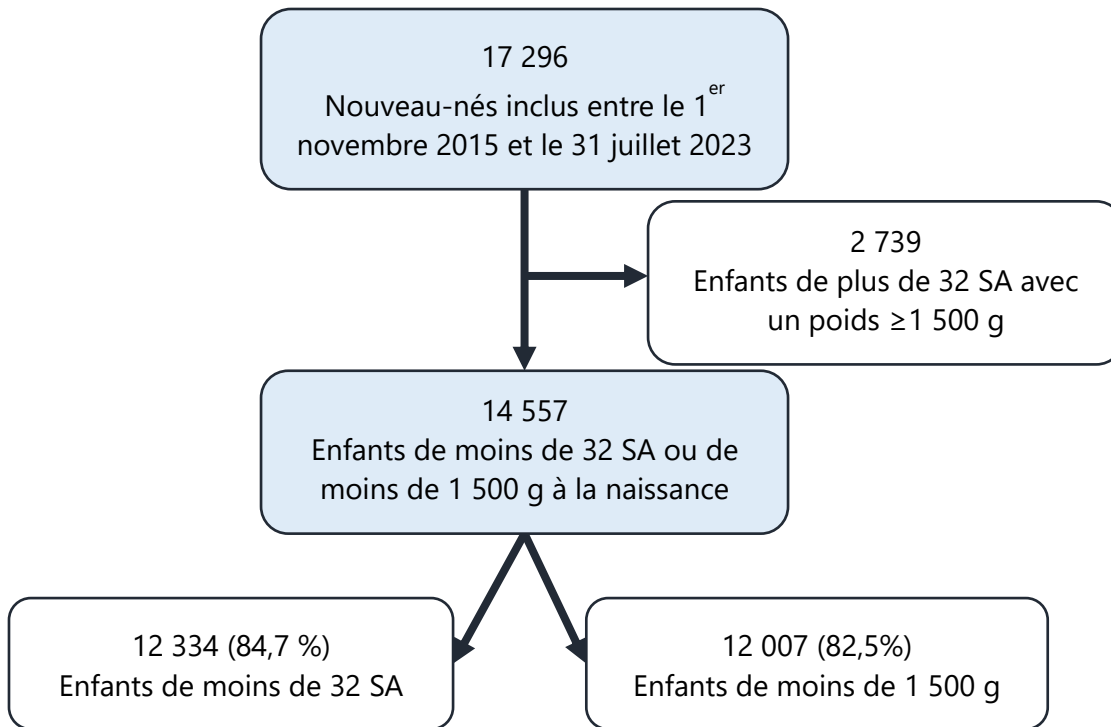


Figure 14 : Sélection de la population d'étude pour la description des données HYGIE-SEV
SA, semaines d'aménorrhée

5.1. Caractéristiques socio-économiques

Les caractéristiques socio-économiques à l'inclusion des nourrissons et de leurs parents sont présentées dans le tableau 6. Une minorité d'enfants (13,1 %) vivaient avec une mère célibataire. Plus de quatre enfants sur cinq (83,5 %) bénéficiaient par leur mère d'une couverture médicale avec sécurité sociale et seulement 15,3 % d'une couverture de type AME ou CMU. Pour la moitié des enfants (50,2 %), la mère avait un niveau d'éducation supérieur (ayant validé au moins une année d'études supérieures après le baccalauréat) et, pour une faible partie (5,3 %), elle n'avait jamais été scolarisée ou seulement en classes primaires. La mère d'un enfant sur cinq (22,8 %) était au chômage ou sans emploi. Une minorité d'enfants (8,2 %) vivaient dans un logement inadéquat (insalubre ou trop petit pour la taille du foyer).

Si 12,3 % vivaient à Paris *intra-muros*, pratiquement les trois quarts 74,3 % avaient leur résidence dans une zone à fort niveau urbain (villes comptant plus de 50 000 habitants). Aucune différence n'était notable sur les caractéristiques socio-économiques entre les enfants de moins de 32 SA et ceux de moins de 1 500 g tant sur la part de données manquantes que sur la répartition des différentes modalités.

Tableau 6 : Caractéristiques socio-économiques des enfants à l'inclusion (N=14 557), cohorte HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023

Caractéristiques	Total (N = 14 557)		< 32 SA (N = 12 334)		< 1 500g (N = 12 007)	
	n	%	n	%	n	%
Année de naissance de l'enfant						
2015-2016	2 393	16,4	1 589	16,7	1 535	16,6
2017-2018	3 786	26,0	3 174	25,8	3 147	26,2
2019-2020	3 764	25,9	3 213	26,0	3 107	25,9
2021-2022	3 633	25,0	3 077	25,0	2 958	24,7
2023	981	6,7	817	6,6	806	6,7
Département de résidence						
Paris (75)	1 792	12,3	1 505	12,2	1 481	12,3
Seine-et-Marne (77)	1 624	11,2	1 386	11,2	1 318	11,0
Yvelines (78)	1 610	11,1	1 336	10,8	1 309	10,9
Essonne (91)	1 707	11,7	1 476	12	1 394	11,6
Hauts-de-Seine (92)	1 593	10,9	1 341	10,9	1 312	10,9
Seine-Saint-Denis (93)	2 640	18,1	2 272	18,4	2 180	18,2
Val-de-Marne (94)	1 715	11,8	1 463	11,9	1 435	12
Val-d'Oise (95)	1 669	11,5	1 381	11,2	1 400	11,7
Hors Ile-de-France	207	1,4	1 505	12,2	178	1,5
Niveau urbain du lieu de vie (nombre d'habitants)						
Moyen ou faible (<50 000)	3 232	25,7	2 618	24,7	2 608	25,2
Elevé (>50 000)	9 325	74,3	7 988	75,3	7 724	74,8
Inconnu*	2 000	13,7	1 728	14,0	1 675	14,0
Age de la mère à la naissance (années)						
<18	47	0,3	41	0,3	36	0,3
18-25	1 815	12,6	1 550	12,7	1 528	12,9
26-32	5 509	38,4	4 709	38,7	4 546	38,3
33-39	4 868	33,9	4 102	33,7	3 997	33,7
≥ 40	2 121	14,8	1 769	14,5	1 751	14,8
Inconnu*	197	1,4	163	1,3	149	1,2
Statut marital de la mère						
En couple	11 529	86,9	9 712	86,6	9 433	86,3
Célibataire	1 738	13,1	1 500	13,4	1 492	13,7
Inconnu*	1 290	8,9	1 122	9,1	1 082	9,0
Couverture sociale						
Sécurité sociale	10 657	83,5	8 929	82,9	8 696	82,9
AME ou CMU	1 954	15,3	1 709	15,9	1 667	15,9
Aucune ou autre	156	1,2	137	1,3	131	1,2
Inconnu*	1 790	12,3	1 559	12,6	1 513	12,6

Tableau 6 (suite)

Caractéristiques	Total (N = 14 557)		< 32 SA (N = 12 334)		< 1 500g (N = 12 007)	
	n	%	n	%	n	%
Conditions de logement						
Adéquates	9 080	75,6	7 658	75,5	7 401	74,9
Inadéquates (logement trop petit ou insalubre)	981	8,2	838	8,3	835	8,5
Chez un proche	889	7,4	741	7,3	758	7,7
Foyer d'accueil	426	3,5	371	3,7	366	3,7
Etablissement d'urgence	137	1,1	120	1,2	118	1,2
Autre	497	4,1	420	4,1	398	4,0
Inconnu*	2 547	17,5	2 186	17,7	2 131	17,7
Nombre de frères et sœurs à la naissance						
0	3 799	32,8	3 179	32,4	3 217	33,7
1	3 653	31,6	3 106	31,6	2 975	31,2
2 ou plus	4 124	35,6	3 535	36,0	3 345	35,1
Inconnu*	2 981	20,5	2 514	20,4	2 470	20,6
Ressource économique du foyer						
Travail d'un parent au moins	11 194	86,8	9 429	86,4	9 168	86,4
Aides sociales	1 167	9,0	1 023	9,4	990	9,3
Autre source	195	1,5	168	1,5	158	1,5
Aucune	343	2,7	295	2,7	296	2,8
Inconnu*	1 658	11,4	1 419	11,5	1 395	11,6
Niveau d'éducation de la mère						
Supérieur	6 441	50,2	5 420	50,1	5 247	49,8
Secondaire (Collège/CAP/Lycée/BAC)	5 699	44,4	4 815	44,5	4 706	44,6
Aucun ou niveau primaire	685	5,3	574	5,3	588	5,6
Inconnu*	1 732	11,9	1 525	12,4	1 466	12,2
Situation socio-professionnelle de la mère						
Travail : cadre et profession intellectuelle	2 226	17,1	1 824	16,5	1 802	16,8
Travail : intermédiaire	754	5,8	648	5,9	623	5,8
Travail : ouvrier ou employé	3 924	30,1	3 264	29,6	3 248	30,2
Travail : profession inconnue	171	1,2	145	1,2	144	1,2
Congé parental	3 157	24,2	2 731	24,8	2 604	24,2
Chômage/Sans emploi	2 971	22,8	2 556	23,2	2 468	23,0
Inconnu*	1 354	9,3	1 166	9,5	1 118	9,3
Situation socio-professionnelle du père						
Travail : cadre et profession intellectuelle	2 646	21,9	2 204	21,5	2 137	21,5
Travail : intermédiaire	1 563	12,9	1 342	13,1	1 264	12,7
Travail : ouvrier ou employé	6 084	50,3	5 143	50,3	5 048	50,7
Travail : profession inconnue	171	1,2	145	1,2	144	1,2
Congé parental	156	1,3	129	1,3	123	1,2
Chômage/Sans emploi	1 646	13,6	1 414	13,8	1 380	13,9
Inconnu*	2 291	15,7	1 957	15,9	1 911	15,9

* Les proportions de données inconnues sont calculées par rapport au nombre total de sujets, tandis que les proportions dans chaque catégorie d'une variable donnée sont calculées par rapport au nombre de sujets ayant des données complètes pour cette variable.

SA, semaines d'aménorrhée ; AME, aide médicale d'Etat ; CMU, couverture médicale universelle

5.2. Caractéristiques de la grossesse

Les caractéristiques de la grossesse et de la période néonatale sont présentées dans le tableau 7. Un peu plus de la moitié (51,4 %) des nourrissons inclus étaient de sexe masculin, et un sur quatre (25,3 %) avait un âge gestationnel de moins de 28 SA. Un peu plus de la moitié (52,9 %) avait un très faible poids de naissance (entre 1 000 g et 1 500 g) et près d'un tiers des enfants (29,4 %) avait un poids de naissance extrêmement faible (moins de 1 000 g). Le poids, la taille et le périmètre crânien moyens (écart-type) à la naissance étaient respectivement de 1 201,1 (336,6) g, 37,5 (3,6) cm et 26,6 (2,5) cm. À la sortie de l'hôpital, le poids, la taille et le périmètre crânien moyens étaient plus élevés, comme attendu, et respectivement de 2 572,9 (549,3) g, 45,5 (3,0) cm et 33,1 (1,9) cm.

Des naissances multiples avaient eu lieu chez un quart (27,0 %) des enfants et une césarienne avait été pratiquée pour plus de deux tiers (67,3 %). Moins d'une mère sur dix rapportait avoir fumé pendant la grossesse (8,1 %). La grande majorité des enfants (84,9 %) ont reçu une corticothérapie anténatale. Plus d'une mère sur 10 (11,9 %) avait souffert de chorioamniotite pendant la grossesse. La naissance prématurée a été induite médicalement pour plus de la moitié des enfants (53,9 %) et près d'un enfant sur 7 (13,8 %) avaient été conçus à l'aide d'une technique de Procréation Médicalement Assistée (PMA).

Les enfants nés grands ou extrêmes prématurés présentaient des moyennes de poids, taille et périmètre crânien supérieures à celles des enfants avec un faible poids de naissance. Les enfants nés grands ou extrêmes prématurés étaient plus souvent de sexe masculin (52,3 % contre 49,1 % chez les enfants de moins de 1 500 g). La survenue de chorioamniotite durant la grossesse était également plus fréquente dans le groupe des grands et extrêmes prématurés (15,0 % contre 12,8 % chez les enfants de moins de 1 500 g). À l'inverse, le taux d'accouchement par césarienne était supérieur dans le groupe des petits poids de naissance (71,0 % contre 63,3 % chez les grands et extrêmes prématurés). Le taux de prématurité provoquée était également plus important (57,9 % contre 48,6 % chez les moins de 32 SA). Aucune différence concernant les autres caractéristiques de la grossesse et de l'accouchement ou la part de données manquantes n'était observée entre les deux groupes.

Tableau 7 : Caractéristiques de la grossesse et de l'accouchement à l'inclusion (N=14 557), cohorte

HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023

Caractéristiques	Total (N = 14 557)		< 32 SA (N = 12 334)		< 1 500g (N = 12 007)	
	n	%	n	%	n	%
Age gestationnel						
A terme (>= 37 SA)	10	0,1	0	0,0	10	0,1
Prématurité moyenne ou tardive (32 à 37 SA)	2 213	15,2	0	0,0	2 213	18,4
Grande prématurité (28 à 32 SA)	8 657	59,5	8 657	70,2	6 114	50,9
Extrême prématurés (<28 SA)	3 677	25,3	3 677	29,8	3 670	30,6
Sexe						
Féminin	7 075	48,6	5 887	47,7	6 116	50,9
Masculin	7 482	51,4	6 447	52,3	5 891	49,1
Poids de naissance (g)						
>=2 500	8	0,1	8	0,1	0	0,0
1 500g à 2 499	2 542	17,5	2 542	20,6	0	0,0
1 000g à 1 499	7 701	52,9	5 573	45,2	7 701	64,1
<1 000	4 306	29,6	4 211	34,1	4 306	35,9
Poids de naissance en g (moyenne, écart-type)	1 201,1	336,6	1 180,9	356,1	1 096,1	261,9
Taille de naissance en cm (moyenne, écart-type)	37,5	3,6	37,2	3,8	36,7	3,2
Périmètre crânien de naissance en cm (moyenne, écart-type)	26,6	2,5	26,3	2,5	26,1	2,4
Poids de sortie en g (moyenne, écart-type)	2 572,9	549,3	2 643,3	550,8	2 542,9	562,1
Taille de sortie en cm (moyenne, écart-type)	45,5	3,0	45,8	3,0	45,2	3,1
Périmètre crânien de sortie en cm (moyenne, écart-type)	33,1	1,9	33,2	2,0	33,0	2,0
Naissance multiple						
Non	10 632	73,0	9 052	73,4	8 807	73,4
Oui	3 923	27,0	3 280	26,6	3 198	26,6
Inconnu*	2	0,0	2	0,0	2	0,0
Tabac durant la grossesse						
Non	11 292	91,9	9 607	92,4	9 307	92
Oui	991	8,1	795	7,6	813	8,0
Inconnu*	2 274	15,6	1 932	15,7	1 887	15,7
Alcool durant la grossesse						
Non	12 066	99,2	10 212	99,1	9 938	99,2
Oui	94	0,8	89	0,9	78	0,8
Inconnu*	2 397	16,5	2 033	16,5	1 991	16,6
Cannabis durant la grossesse						
Non	12 018	99,2	10 180	99,2	9 903	99,2
Oui	93	0,8	78	0,8	79	0,8
Inconnu*	2 446	16,8	2 076	16,8	2 025	16,9
Autre produit toxique durant la grossesse						
Non	11 764	99,8	9 965	99,7	9 687	99,7
Oui	29	0,2	25	0,3	25	0,3
Inconnu*	2 764	19,0	2 344	19,0	2 295	19,1
Produit de substitution durant la grossesse						
Non	11 902	99,8	10 084	99,8	9 810	99,8
Oui	26	0,2	25	0,2	22	0,2
Inconnu*	2 629	18,1	2 225	18,0	2 175	18,1
Corticostéroïdes anténataux						
Non	1 957	15,1	1 613	14,7	1 539	14,4
Oui	10 977	84,9	9 347	85,3	9 133	85,6
Inconnu*	1 623	11,1	1 374	11,1	1 335	11,1

Tableau 7 (suite)

Caractéristiques	Total (N = 14 557)		< 32 SA (N = 12 334)		< 1 500g (N = 12 007)	
	n	%	n	%	n	%
Chorioamniotite						
Non	10 248	87,1	8 450	85,0	8 481	87,2
Oui	1 516	12,9	1 491	15,0	1 245	12,8
Inconnu*	2 793	19,2	2 393	19,4	2 281	19,0
Mode d'accouchement						
Voie basse	4 595	32,7	4 356	36,7	3 359	29,0
Césarienne	9 443	67,3	7 522	63,3	8 229	71,0
Inconnu*	519	3,6	456	3,7	419	3,5
Prématurité provoquée						
Non	5 780	46,1	5 443	51,4	4 356	42,1
Oui	6 755	53,9	5 148	48,6	6 003	57,9
Inconnu*	2 022	13,9	1 743	14,1	1 648	13,7
Procréation médicalement assistée						
Non	9 783	86,2	8 278	86,4	8 052	86,1
Oui	1 571	13,8	1 306	13,6	1 305	13,9
Inconnu*	3 203	22,0	2 750	22,3	2 650	22,1

* Les proportions de données inconnues sont calculées par rapport au nombre total de sujets, tandis que les proportions dans chaque catégorie d'une variable donnée sont calculées par rapport au nombre de sujets ayant des données complètes pour cette variable.

SA, semaines d'aménorrhée

5.3. Caractéristiques néonatales

La durée de l'hospitalisation initiale était supérieure à 70 jours pour près d'un tiers des enfants (33,0 %, tableau 8). Durant cette période, une minorité de nourrissons ont eu un examen auditif nécessitant une surveillance (8,6 %) ou une rétinopathie de stade II ou plus (3,0 %), ont subi une intervention chirurgicale à la naissance (10,6 %) ou ont reçu des corticostéroïdes postnataux (14,0 %). Plus de la moitié en revanche ont eu une prescription d'anticorps monoclonaux anti-VRS (59,7 %). Plus d'un enfant sur cinq (21,4 %) présentait une dysplasie broncho-pulmonaire et une faible proportion seulement (6,2 %) d'autres pathologies fœtales telles que des malformations ou des syndromes génétiques. Les évaluations neurologiques ont été classées comme présentant un risque modéré (6,8 %) ou élevé (1,9 %) pour une minorité d'enfants. Plus de la moitié des enfants (52,2 %) étaient allaités à leur sortie d'hôpital.

Aucune différence particulière n'était observée entre les deux groupes si ce n'est la part des hospitalisations de courte durée (0 à 30 jours) qui était plus importante chez les enfants de petit poids de naissance (6,2 % contre 2,5 % chez les moins de 32 SA) tandis que les hospitalisations de durée moyenne (31 à 50 jours) étaient plus fréquentes chez les grands ou

extrêmes prématurés (28,4 % contre 25,7 % chez les très faibles poids de naissance). Enfin, le taux de prescription d'un anti-VRS était plus important chez les moins de 32 SA que chez les moins de 1 500 g (respectivement 67,8 % et 62,4 %).

5.1. Caractéristiques de prise en charge médicale

Les caractéristiques du suivi sont présentées dans le tableau 9. La majorité (83,2 %) des enfants étaient nés dans une maternité de niveau 3 (risque le plus élevé), tandis que moins de la moitié (41,6 %) ont été transférés depuis leur maternité de naissance. Ces nourrissons ont été inclus à leur sortie dans les six RSP de la région Ile-de-France dans des proportions comparables, comprises entre 10,0 % pour le réseau RPVM et 27,2 % pour le réseau RPSOF. A la sortie de l'hôpital, le suivi a été assuré par un médecin hospitalier pour deux tiers (66,4 %) des enfants et par un médecin libéral pour un quart (26,0 %). Au moment de l'inclusion, plus d'un tiers des enfants (38,1 %) habitaient à une distance inférieure à 3 kilomètres de leur médecin pilote, tandis que moins d'un sur cinq (17,1 %) vivaient à une distance supérieure à 10 kilomètres.

Selon les caractéristiques de prise en charge médicale à l'inclusion, il n'y avait pas de différence entre les deux groupes si ce n'est un taux légèrement plus important chez les enfants grands ou extrêmes prématurés de transfert de maternité entre la naissance et l'inclusion (42,6 % d'enfants transférés contre 40,6% dans l'autre groupe).

Tableau 8 : Caractéristiques néonatale à l'inclusion (N=14 557), cohorte HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023

Caractéristiques	Total (N = 14 557)		< 32 SA (N = 12 334)		< 1 500g (N = 12 007)	
	n	%	n	%	n	%
Délai d'hospitalisation (jours)						
0-30	856	6,4	289	2,5	691	6,2
31-50	4 391	32,7	3 234	28,4	2 853	25,7
51-70	3 737	27,9	3 510	30,9	3 255	29,3
> 70	4 434	33,0	4 342	38,2	4 296	38,7
Inconnu*	1 139	7,8	959	7,8	912	7,6
Résultats auditifs						
Normal	11 740	87,1	9 880	86,7	9 612	86,4
A contrôler	1 160	8,6	1 004	8,8	1 019	9,2
Examen non fait	579	4,3	518	4,5	488	4,4
Inconnu*	1 078	7,4	932	7,6	888	7,4
Rétinopathie de stade 2 ou plus						
Non	14 122	97,0	11 902	96,5	11 574	96,4
Oui	435	3,0	432	3,5	433	3,6
Corticostéroïdes postnataux						
Non	9 901	86,0	8 175	83,9	7 968	83,7
Oui	1 606	14,0	1 572	16,1	1 548	16,3
Inconnu*	3 050	21,0	2 587	21,0	2 491	20,7
Traitement anti-virus respiratoire syncytial						
Non	5 859	40,3	3 971	32,2	4 504	37,6
Oui	8 677	59,7	8 344	67,8	7 482	62,4
Inconnu*	21	0,1	19	0,2	21	0,2
Intervention chirurgicale à la naissance						
Non	10 559	89,4	8 828	88,4	8 580	88,1
Oui	1 254	10,6	1 159	11,6	1 164	11,9
Inconnu*	2 744	18,9	2 347	19,0	2 263	18,8
Dysplasie bronchopulmonaire						
Non	11 417	78,6	9 250	75,1	9 003	75,1
Oui	3 116	21,4	3 062	24,9	2 980	24,9
Inconnu*	24	0,2	22	0,2	24	0,2
Pathologie fœtale additionnelle						
Non	13 650	93,8	11 602	94,1	11 221	93,5
Oui	901	6,2	726	5,9	781	6,5
Inconnu*	6	0,0	6	0,0	5	0,0
Évaluation neurologique						
Pas de risque	949	6,5	764	6,2	791	6,6
Risques faibles	12 341	84,8	10 398	84,3	10 100	84,1
Risques modérés	996	6,8	921	7,5	878	7,3
Risques élevés	271	1,9	251	2,0	238	2,0
Allaitement prévu à la sortie						
Oui	5 519	47,8	4 742	48,7	4 674	49,2
Non	6 039	52,2	5 005	51,3	4 829	50,8
Inconnu*	2 999	20,6	2 587	21,0	2 504	20,9

* Les proportions de données inconnues sont calculées par rapport au nombre total de sujets, tandis que les proportions dans chaque catégorie d'une variable donnée sont calculées par rapport au nombre de sujets ayant des données complètes pour cette variable.

SA, semaines d'aménorrhée

Tableau 9 : Caractéristiques de la prise en charge à l'inclusion (N=14 557), cohorte HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023

Caractéristiques	Total (N = 14 557)		< 32 SA (N = 12 334)		< 1 500g (N = 12 007)	
	n	%	n	%	n	%
Réseau d'inclusion						
RPSOF – Région Sud-Ouest	3 956	27,2	3 364	27,3	3 216	26,8
RSPP – Paris	2 843	19,5	2 445	19,8	2 439	20,3
NEF – Région Nord-Est	3 144	21,6	2 684	21,8	2 541	21,2
RPVM – Région Sud-Est	1 461	10,0	1 239	10,0	1 214	10,1
RPVO – Région Nord-Ouest	1 675	11,5	1 380	11,2	1 399	11,7
MYP A – Région Ouest	1 478	10,2	1 222	9,9	1 198	10,0
Médecin pilote						
Hospitalier (public/privé)	8 573	66,4	7 565	68,9	7 352	68,9
Libéral	3 355	26,0	2 645	24,1	2 566	24,1
Autre	990	7,7	768	7,0	751	7,0
Inconnu*	1 639	11,3	1 356	11,0	1 338	11,1
Type de maternité						
Niveau 3	11 475	83,2	9 868	84,7	9 653	84,7
Niveau 1 ou 2	2 323	16,8	1 784	15,3	1 743	15,3
Inconnu*	759	5,2	682	5,5	611	5,1
Enfant né en dehors de son établissement d'inclusion						
Non	8 166	58,4	6 794	57,4	6 854	59,4
Oui	5 814	41,6	5 032	42,6	4 677	40,6
Inconnu*	577	4,0	508	4,1	476	4,0
Distance entre le lieu de domicile et le médecin pilote (km)						
<3	4 561	38,9	3 783	37,9	3 688	37,9
3-10	5 162	44,0	4 420	44,3	4 335	44,5
>10	2 001	17,1	1 779	17,8	1 715	17,6
Inconnu*	2 833	19,5	2 352	19,1	2 269	18,9

* Les proportions de données inconnues sont calculées par rapport au nombre total de sujets, tandis que les proportions dans chaque catégorie d'une variable donnée sont calculées par rapport au nombre de sujets ayant des données complètes pour cette variable.

SA, semaines d'aménorrhée

6. Forces et limites du programme

La cohorte HYGIE-SEV se distingue par sa conception prospective et ouverte, qui permet de recueillir des données objectives sur le développement neurologique et la croissance d'un grand nombre d'enfants, de la naissance à l'âge de sept ans, depuis 2015. Au 1^{er} juillet 2023, la cohorte comptait plus de 14 000 enfants, ce qui lui confère la puissance statistique suffisante pour étudier un large éventail d'hypothèses sur la santé et les soins périnataux dans diverses sous-populations. La cohorte est basée sur la population générale, en région parisienne. Cette région couvrant près de 20 % de la population française (>12 millions de personnes) sur un territoire aux conditions socio-économiques inéquitables (169) et aux conditions démographiques singulières de par l'attraction migratoire intra-territoire

concernant la prise en charge de l'accouchement (170,171) et l'importance du suivi dans le cadre de la périnatalité (172), la cohorte HYGIE-SEV est rendue plus représentative de la réalité qu'une cohorte de recherche traditionnelle.

Il est désormais reconnu qu'une grande variété de conditions influence les résultats et les soins de santé des enfants nés très prématurément ou avec un très faible poids de naissance (173,174). La base de données HYGIE-SEV contient plus d'une centaine de variables décrivant un large éventail de conditions telles que les complications prénatales, les caractéristiques et morbidités périnatales, les conditions socio-économiques ainsi que l'organisation et l'efficacité du système de soins de santé. Cette richesse de variables permettra d'étudier l'influence d'une multitude de facteurs, de développer des études prenant en compte des facteurs multidimensionnels et de mieux quantifier leur impact relatif sur la trajectoire de santé des enfants.

HYGIE-SEV offre des perspectives encore inexplorées par les autres cohortes internationales (cf. tableau S1 en annexe 1) non seulement en ce qui concerne les données collectées mais aussi au regard de la population incluse. Ainsi, les enfants grands prématurés inclus dans ce programme représentent un registre régional avec une exhaustivité de plus de 80 % du bassin de recrutement dont ils font partie (172). Alors que la plupart des cohortes existantes se sont intéressées exclusivement au devenir des enfants prématurés, la grande prématurité n'était pas le seul critère de vulnérabilité parmi les critères d'inclusion du programme HYGIE-SEV. Nous avons vu que les enfants de très faible poids de naissance sont aussi inclus dans HYGIE-SEV et représentent une population d'effectif quasiment équivalent à celle des grands ou extrêmes prématurés, ce qui permet des analyses statistiques spécifiques sur ce groupe également. D'autres enfants sont également suivis dans le programme HYGIE-SEV selon d'autres critères de vulnérabilité que les enfants grands prématurés et de très faible poids de naissance. Ils représentent cependant un groupe hétérogène et d'effectif encore trop limité (2 739 enfants, soit 15,8 % des enfants inclus entre le 1^{er} novembre 2015 et le 31 juillet 2023, figure 14) pour être étudié spécifiquement dans ces travaux de thèse. Ils n'en demeurent pas moins une source de données exploitables pour des études ultérieures.

Une des limites de la cohorte est l'utilisation de scores non validés pour évaluer les différents aspects de la santé des enfants. Bien qu'il existe de nombreux tests standardisés permettant d'évaluer l'état de santé des enfants prématurés, tels que l'échelle d'intelligence préscolaire

et primaire de Wechsler (WPPSI) (175) pour déterminer le quotient intellectuel ou l'échelle de Bayley pour le développement des nourrissons et des enfants en bas âge (*Bayley Scale of Infant and Toddler Development-III*) (176), ces tests peuvent prendre beaucoup de temps et nécessiter de nombreuses ressources ; c'est pourquoi ils ne sont pas utilisés dans la cohorte. En revanche, de futures études de validation seront menées pour confirmer la fiabilité des scores utilisés dans la cohorte HYGIE-SEV.

Bien qu'un grand nombre de variables soient collectées dans la cohorte HYGIE-SEV, certaines informations manquent, en particulier en ce qui concerne la période de grossesse. À l'avenir, le comité scientifique prévoit d'augmenter les données de la cohorte en les chaînant à celles du Système National des Données de Santé (SNDS) (177). Cela permettra d'obtenir des informations supplémentaires sur les nourrissons, leur mère et les soins de santé associés pendant l'enfance. De plus, d'autres types de variables qualifiant le niveau socio-économique du lieu de vie n'ont pas été pris en compte dans les analyses présentées ici ou dans la suite de la thèse. Par la suite, l'indice de désavantage social (FDep) ou encore le pourcentage d'étrangers pourront par exemple être ajoutés aux variables déjà disponibles.

L'absence de groupe « témoin » d'enfants non-prématurés peut également être une limite car la comparaison directe entre cette population témoin et la population cas ne peut pas être effectuée.

La dernière limite est la possibilité d'une inclusion non exhaustive, où certains sous-groupes de la population cible peuvent ne pas avoir été représentés de manière adéquate. Le groupe des enfants grands prématurés dans la cohorte HYGIE-SEV représentait une exhaustivité de 80 % (172) mais une analyse plus approfondie est nécessaire pour quantifier pleinement l'exhaustivité des inclusions dans l'ensemble de la cohorte afin de prendre en compte tous les biais potentiels qui peuvent en résulter.

CHAPITRE 3 –

OPTIMALITE DU SUIVI : ANALYSE DES FACTEURS ASSOCIES

Ce troisième chapitre présente les résultats d'un deuxième travail de thèse s'intéressant à l'optimalité du suivi dans le programme HYGIE-SEV. Il vise en particulier à trouver des facteurs associés à la perte de vue afin de déterminer des premières pistes d'explication en lien avec l'attrition dans le but d'augmenter la rétention dans le programme HYGIE-SEV. Dans ce chapitre comme dans l'intégralité de cette thèse, les perdus de vue sont considérés comme étant des « non-réponses » soit les enfants sortis du dispositif HYGIE-SEV.

1. Introduction

1.1. Problématique générale

A cause du suivi lourd et fastidieux pour les enfants, pour les familles et aussi pour les médecins, la non-adhésion est inévitable dans un suivi médical à long terme des enfants vulnérables. Ainsi, un grand nombre de ces enfants est généralement perdu de vue (66,74,75). Ces perdus de vue représentent une occasion manquée de détection et de traitement précoces des séquelles éventuelles causées par la prématurité et autres facteurs de vulnérabilité. L'identification des facteurs associés à l'attrition est essentielle pour concevoir des interventions efficaces de rétention des patients dans cette population. Pour comprendre les raisons de l'attrition élevée dans les cohortes de prématurés, plusieurs études ont tenté d'analyser les facteurs de risque de perte de suivi (71–73,79,80,178–181). La plupart de ces études se sont concentrées sur les résultats à court terme, généralement vers l'âge de deux ans, et se sont limitées à une seule dimension de variables explicatives, socioéconomiques et démographiques (66,71–73,79,80,178,180,181), cliniques (66,73,79,80,178–180) ou

géographiques (66,79,80), ou encore des facteurs spécifiques à la cohorte ou au programme (73,80,179). Cette hétérogénéité méthodologique des études publiées aboutit à des résultats dispersés (74,75).

1.2. Problématiques liées au suivi en Île-de-France

Le suivi des enfants vulnérables en Île-de-France est complexe du fait des particularités uniques de cette zone géographique par rapport aux autres régions françaises métropolitaines. Ces particularités sont les suivantes :

1. L'Île-de-France est la région la plus densément peuplée de France, comptant 12,4 millions d'habitants (7), ce qui équivaut à la population d'un pays comme la Belgique (8).
2. La région connaît un nombre important de naissances, avec 161 200 naissances enregistrées en 2022 selon l'INSEE (171).
3. Les inégalités sociales y sont particulièrement prononcées (169). Cette disparité socio-économique crée des défis supplémentaires pour assurer un suivi adéquat à l'ensemble des enfants vulnérables.
4. L'Île-de-France enregistre le plus grand nombre de mouvements migratoires en France, tant vers ou depuis d'autres régions qu'entre les départements qui la composent (170). Ces déménagements fréquents posent des défis spécifiques en matière de gestion du suivi des enfants vulnérables, nécessitant une coordination efficace entre les différents services et réseaux.
5. L'organisation du suivi des enfants vulnérables repose sur plusieurs réseaux de suivi, ce qui entraîne des problématiques de gestion du suivi, en particulier dans un contexte régional où les déménagements sont fréquents.
6. L'Île-de-France est une destination attractive pour les accouchements. Il y a plus de naissances et d'accouchements dans les maternités de la région que de naissances et d'accouchements des résidents de l'Île-de-France eux-mêmes (171). Cette situation ajoute une pression supplémentaire sur le réseau de suivi des enfants vulnérables (172).

La prise en charge des enfants vulnérables dans la région Ile-de-France est donc contraint par de multiples facteurs qui peuvent déboucher sur des difficultés de suivi. Ainsi, lorsqu'on s'intéresse à l'évolution du suivi dans le programme HYGIE-SEV, on s'aperçoit que le taux de visites non réalisées augmente au fur et à mesure que le suivi avance (figure 15). Ce taux est de 39,0 % pour la visite des deux ans et atteint 62,6 % à cinq ans. A partir de la visite des quatre ans, le taux de visites non réalisées devient plus important que le taux de visites réalisées. Etant donné le caractère continu des inclusions dans le programme HYGIE-SEV, ces taux sont calculés en rapport le nombre de visites réalisées au nombre d'enfants éligibles à la visite, c'est-à-dire les enfants ayant atteint l'âge de la visite + 6 mois (cf. section 2.1).

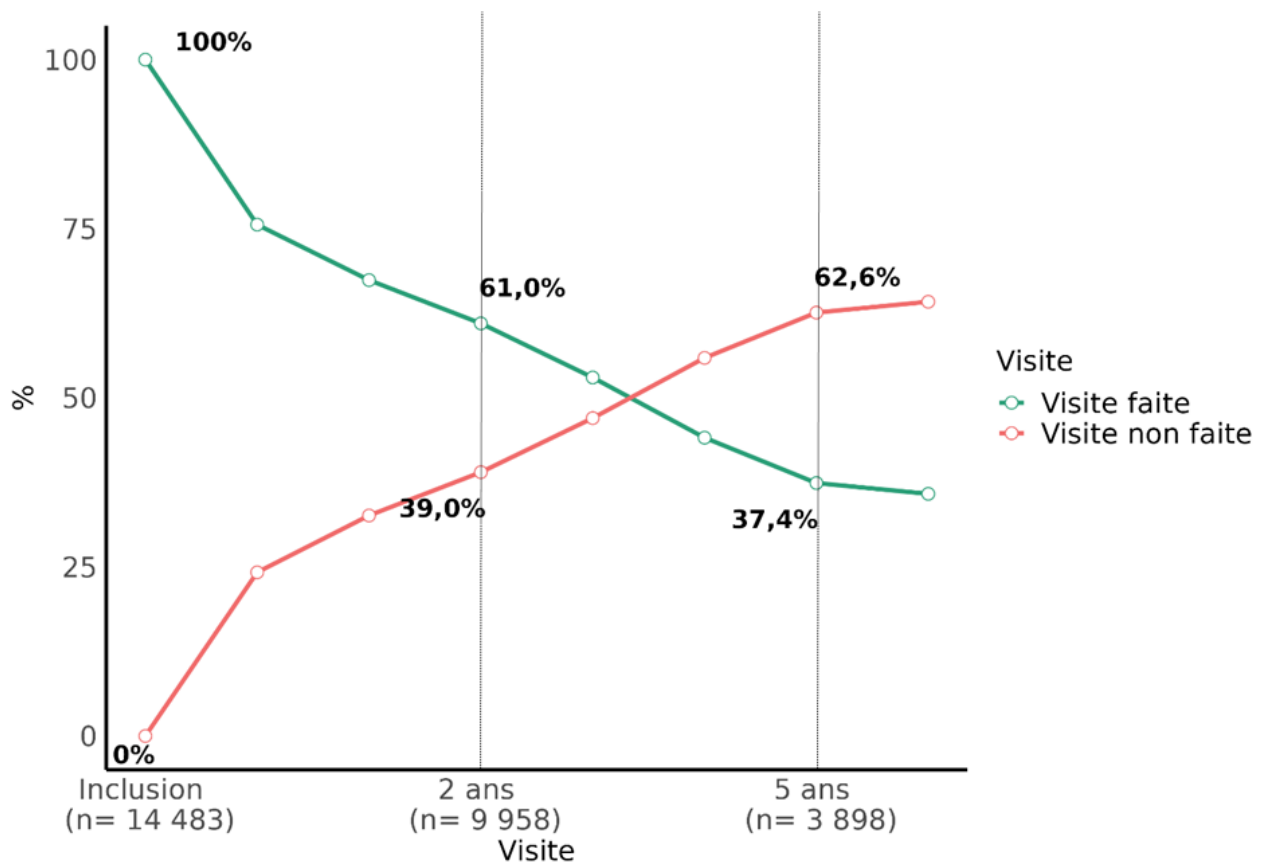


Figure 15 : Evolution des proportions de visites réalisées (taux de suivi, en vert) et non réalisées (taux d'attrition, en rouge) dans le programme HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023

2. Méthodes

2.1. Population incluse

Pour cette étude, nous nous sommes limités à la population des grands prématurés selon les critères d'inclusion suivants : enfants nés entre le 1^{er} novembre 2015 et 31 octobre 2022 (soit six années complètes) avec un âge gestationnel inférieur à 33 semaines d'aménorrhée.

Les résultats d'intérêt pour la présente étude étaient les taux d'attrition aux visites obligatoires à deux et cinq ans dans le programme HYGIE-SEV. Un délai de 6 mois au-delà de la date anniversaire était autorisé. Les enfants éligibles à la visite des deux ans étaient donc ceux qui avaient plus de 2,5 ans d'âge corrigé à la date d'extraction des données, c'est-à-dire ceux dont la date de naissance prévue était antérieure au 1^{er} mai 2020. De même, les enfants éligibles à la visite des cinq ans étaient ceux qui avaient plus de 5,5 ans à la date d'extraction des données, c'est-à-dire ceux qui étaient nés avant le 1^{er} mai 2017. Les enfants décédés ou ayant quitté la région Ile-de-France avant l'âge de deux ou cinq ans ont été exclus des analyses correspondantes (figure 16).

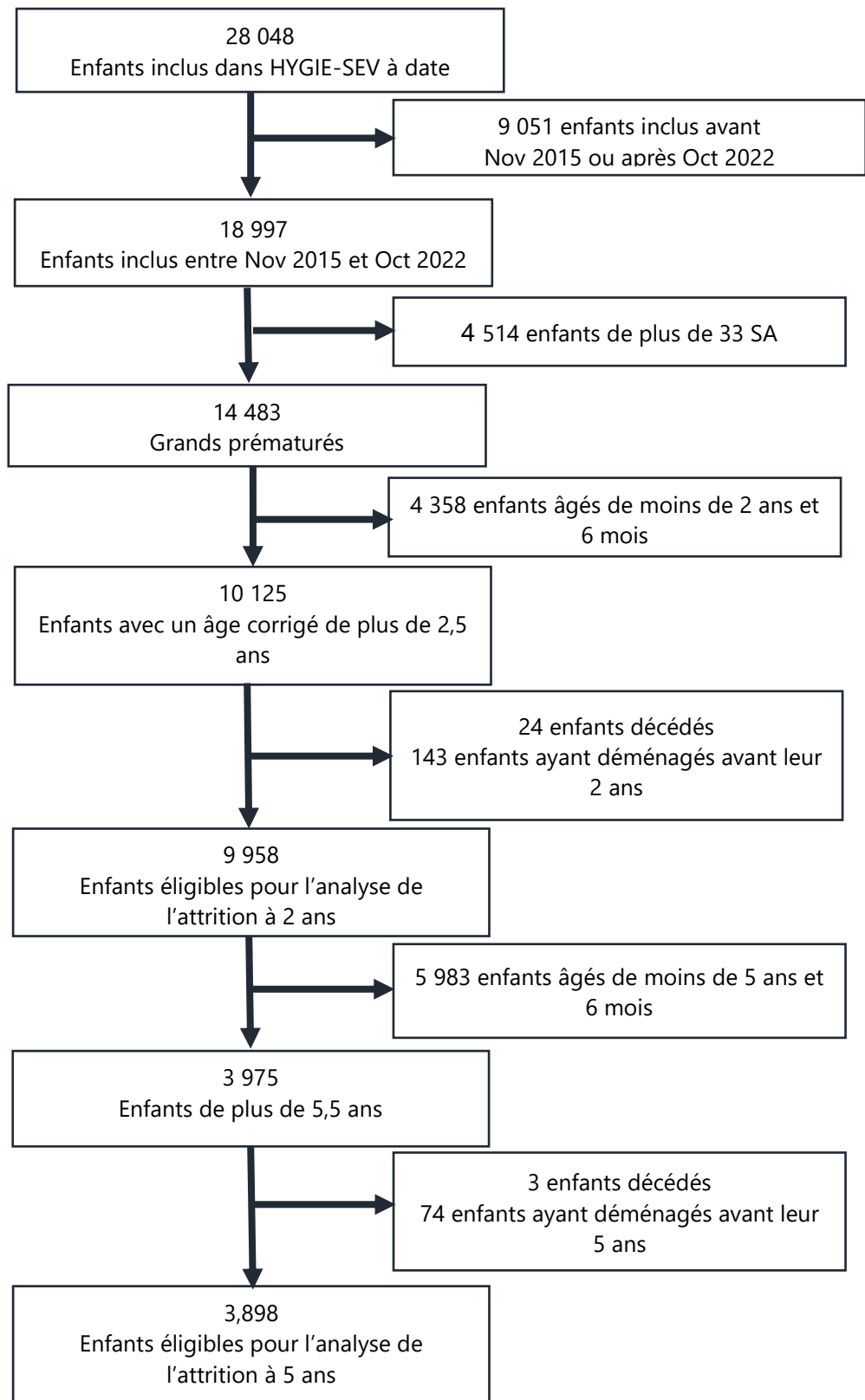


Figure 16 : Diagramme de flux des enfants inclus dans les analyses à deux et cinq ans, cohorte HYGIE-SEV, 1er novembre 2015 - 31 juillet 2023

2.2. Variables d'exposition

Nous avons évalué simultanément plusieurs dimensions comme facteurs de risque potentiels de perte de suivi : les caractéristiques sociodémographiques des parents, les caractéristiques médicales de la mère et de sa grossesse, les caractéristiques périnatales de l'enfant, ainsi que des variables contextuelles évaluées lors de l'inscription au programme HYGIE-SEV.

Les caractéristiques sociodémographiques parentales étaient : le niveau d'éducation et le statut marital de la mère, les ressources économiques du foyer, la couverture médicale, la situation professionnelle et la profession de chaque parent, les conditions de logement et le nombre d'enfants dans le foyer. Les conditions de logement inadéquates ont été définies comme un logement individuel trop petit pour la taille du ménage ou dépourvu des commodités essentielles telles que l'assainissement, l'eau et le chauffage.

Parmi les caractéristiques médicales de la mère et de sa grossesse, nous avons pris en compte l'âge maternel à la naissance, une grossesse multiple, le mode d'accouchement, l'administration anténatale de corticostéroïdes, la survenue d'une chorioamniotite, le tabagisme et la consommation d'alcool, de cannabis, d'autres produits toxiques ou de produits de substitution pendant la grossesse.

Les caractéristiques périnatales comprenaient le sexe de l'enfant, l'âge gestationnel, la durée du séjour avant la sortie de l'hôpital, les Z-scores du poids et du périmètre crânien à la naissance, les différences (appelées delta) dans les Z-scores du poids de naissance et du poids à la sortie de l'hôpital, l'année de naissance, le recours à l'assistance médicale à la procréation, la prématurité induite, la naissance dans un hôpital différent de la maternité d'inclusion, les résultats de l'examen de l'audition, la rétinopathie du prématuré (stade II ou supérieur), la dysplasie broncho-pulmonaire, la prescription postnatale de corticostéroïdes systémiques ou inhalés, la prescription d'anticorps monoclonaux anti-Virus Respiratoire Syncytial (VRS), une pathologie foetale additionnelle, le niveau de risque d'exploration neurologique, une intervention chirurgicale à la naissance et l'allaitement maternel à la sortie de l'hôpital. L'âge gestationnel a été calculé à partir de la date prévue de l'accouchement, estimée à partir de la date des dernières règles de la mère en SA. Les Z-scores du poids et du périmètre crânien pour un âge donné ont été calculés selon les normes de Fenton et al. (182). La dysplasie

broncho-pulmonaire a été définie comme le besoin d'oxygène supplémentaire ou le besoin d'une assistance respiratoire à 36 SA.

En plus des variables individuelles recueillies par les médecins, nous avons pris en compte les variables contextuelles suivantes décrivant le contexte géographique et médical à l'inclusion : le niveau urbain du lieu de résidence au moment de la naissance telle que définie par l'INSEE (165), la distance entre le domicile et le médecin pilote, le RSP d'inclusion, le niveau de la maternité de naissance et le type de médecin pilote (hospitalier ou libéral).

2.3. Définition du critère de jugement

Un enfant a été classé comme perdu de vue lors de la visite à deux ou cinq ans s'il n'avait pas effectué la visite correspondante dans les 6 mois suivant la date prévue, indépendamment de toute visite antérieure ou ultérieure. Initialement d'autres définitions de perdu de vue avaient été envisagées, telle que la non présentation à deux visites consécutives sans être revu dans les visites ultérieures. Cette définition, bien que possiblement satisfaisante d'un point de vue médical, ne l'était pas d'un point de vue méthodologique. En effet, la définition de la perte de vue basée sur des critères évalués a posteriori pose un problème de temporalité. Ainsi, un enfant de moins de trois ans serait forcément qualifié de non perdu de vue à deux ans car il n'aurait pas eu le temps d'effectuer sa visite des trois ans et il n'aurait donc pas été possible de déterminer s'il aurait manqué deux visites consécutives. De même, un enfant de quatre ans ayant manqué les visites des deux ans et des trois ans serait identifié comme perdu de vue sans savoir s'il reviendrait à des visites ultérieures. Une solution aurait consisté à ne sélectionner que les enfants ayant terminé leur suivi obligatoire (c'est-à-dire les enfants ayant atteint sept ans et plus à la date d'extraction des données) afin de se baser sur un échantillon homogène en termes d'âge. Cependant, au moment d'aborder les questions de recherche sur les perdus de vue, nous ne disposons pas d'un échantillon suffisamment grand pour fournir des analyses représentatives. C'est pourquoi nous sommes revenus en arrière et avons opté pour une définition plus pragmatique mais aussi moins biaisée. Toutefois, dans une analyse de sensibilité, nous avons également envisagé une définition plus stricte : parmi les enfants éligibles à la visite des cinq ans, seuls les enfants qui ne se sont pas présentés à la

visite des deux ans et qui n'ont pas été revus avant la visite des cinq ans ont été classés comme perdus de vue à la visite des deux ans.

2.4. Analyses statistiques

Les distributions des variables ont été décrites à l'aide d'effectifs et de proportions. Une imputation multiple (IM) des données manquantes parmi les variables d'exposition a été mise en œuvre en utilisant toutes les variables, selon un algorithme *bootstrap* et de maximisation de l'espérance tel qu'implémenté dans le package MICE de R (183). Cinquante ensembles de données imputées indépendantes ont été générés. Nous avons utilisé des modèles de régression logistique univariés pour évaluer les associations brutes des caractéristiques étudiées avec la non-participation à la visite des deux ans ou des cinq ans avant et après IM. Les associations ont été exprimées sous forme d'odds ratio (OR) et d'intervalles de confiance à 95 % (IC95%).

Pour sélectionner les variables dans notre modèle multivarié final, seuls les facteurs ayant une p global (test du rapport de vraisemblance) $< 0,2$ dans les modèles univariés ont été conservés, à l'exception du sexe, de l'âge gestationnel, du réseau d'inclusion et de l'année de naissance qui ont été forcés dans tous les modèles. Les variables dont le facteur d'inflation VIF était supérieur à 5 ont été exclues (184). La valeur du facteur d'inflation de la variable, ou *Variance Inflation Factor* (VIF) (185), est une mesure de la multicolinéarité du modèle. La multicolinéarité existe lorsque plusieurs facteurs supposés indépendants dans le modèle de régression sont corrélés (cf. chapitre 1). Cela peut affecter l'estimation des coefficients de régression associés à ces facteurs. Ainsi le VIF permet d'estimer de combien la variance d'un coefficient de régression est augmentée du fait de la multicolinéarité. La valeur du VIF est calculée pour chaque variable explicative par la formule suivante : $1/(1-R^2)$ ou R^2 représente le coefficient de détermination de la régression de la variable explicative considérée sur toutes les autres. La durée d'hospitalisation a également été exclue car elle était fortement corrélée à l'âge gestationnel (coefficient de corrélation de Spearman = 0,74).

Toutes les analyses ont été effectuées à l'aide du logiciel statistique R 4.1.1. Tous les tests statistiques étaient bilatéraux. Les valeurs de p inférieures à 0,05 ont été considérées comme statistiquement significatives.

3. Résultats

3.1. Analyses descriptives

Entre le 1^{er} novembre 2015 et le 31 octobre 2022, 14 483 prématurés dont l'âge gestationnel était inférieur à 33 SA ont été inclus dans le programme HYGIE-SEV. Parmi eux, 9 958 étaient éligibles à la visite des deux ans et 3 898 à la visite des cinq ans (figure 16). Parmi les nourrissons éligibles, 3 881 (39,0 %) ont manqué la visite à deux ans et 2 440 (62,6 %) la visite à cinq ans. Selon la définition plus stricte (analyse de sensibilité), 1 222 (31,3 %) ont été considérés comme perdus de vue à la visite des deux ans.

Les caractéristiques de tous les enfants inclus et de ceux qui étaient éligibles aux visites des deux et cinq ans sont présentées dans le tableau 10. Les distributions étaient globalement similaires dans les trois populations. On peut cependant noter quelques variations : en particulier, le taux de données manquantes tendait à baisser à mesure que la sélection de patients se basait sur des enfants plus âgés pour la plupart des variables à l'exception de la couverture sociale, des conditions de logement, du nombre d'enfants présents dans le foyer, le mode d'accouchement, le poids à la sortie, la prématurité provoquée, la procréation médicalement assistée, les corticostéroïdes anténataux, la chorioamniotite, les résultats auditifs, les corticostéroïdes postnataux, la durée d'hospitalisation, le type de médecin pilote et le niveau de maternité de naissance. Les taux de non-réponse variaient selon les variables et se situaient le plus souvent entre 0 et 25 %.

Tableau 10 : Caractéristiques des patients à l'inclusion (N=14 483), à la visite à deux ans (N=9 958) et à la visite à cinq ans (N=3 898), cohorte HYGIE-SEV, 1er novembre 2015 - 31 octobre 2022

Caractéristique	Inclusion (N=14 483)		2 ans (âge corrigé) (N=9 958)		5 ans (N=3 898)	
	n	%	n	%	n	%
Niveau d'éducation de la mère						
Supérieur	6 436	50,2	4 394	49,6	1 766	50,3
Secondaire (Collège/CAP/Lycée/BAC)	5 730	44,7	4 000	45,1	1 569	44,7
Aucun ou niveau primaire	650	5,1	466	5,3	176	5,0
Inconnu*	1 667	11,5	1 098	11,0	387	9,9
Ressource économique du foyer						
Travail d'un parent au moins	11 168	86,8	7 680	86,4	3 048	86,7
Aides sociales	1 184	9,2	836	9,4	313	8,9
Autre source	194	1,5	142	1,6	65	1,8
Aucune	322	2,5	226	2,5	89	2,5
Inconnu*	1 441	10,1	971	9,9	349	9,0
Situation socio-professionnelle de la mère						
Travail : cadre et profession intellectuelle	2 241	16,9	1 515	16,5	580	15,9
Travail : intermédiaire	792	6,0	507	5,5	213	5,9
Travail : ouvrier ou employé	3 820	28,8	2 655	28,9	1 075	29,6
Travail : profession inconnue	327	2,5	251	2,7	132	3,6
Congé parental	3 163	23,8	2 232	24,3	899	24,7
Chômage/Sans emploi	2 938	22,1	2 023	22,0	738	20,3
Inconnu*	1 202	8,3	775	7,8	261	6,7
Situation socio-professionnelle du père						
Travail : cadre et profession intellectuelle	2 733	21,8	1 839	21,2	735	21,4
Travail : intermédiaire	1 606	12,8	1 074	12,4	396	11,5
Travail : ouvrier ou employé	5 996	47,7	4 184	48,2	1 629	47,5
Travail : Profession inconnue	476	3,8	374	4,3	198	5,8
Congé parental	139	1,1	82	0,9	32	0,9
Chômage/Sans emploi	1 612	12,8	1 125	13,0	441	12,9
Inconnu*	1 921	13,3	1 280	12,9	467	12,0
Couverture sociale						
Sécurité sociale	10 608	83,6	7 257	83,7	2 770	84,4
AME ou CMU	1 929	15,2	1 301	15,0	468	14,3
Aucune ou autre	156	1,2	114	1,3	43	1,3
Inconnu*	1 790	12,4	1 286	12,9	617	15,8
Conditions de logement						
Adéquates	9 157	76,1	6 413	76,0	2 578	76,5
Inadéquates (logement trop petit ou insalubre)	983	8,2	726	8,6	304	9,0
Chez un proche	882	7,3	648	7,7	262	7,8
Foyer d'accueil, établissement d'urgence ou autre	1 009	8,4	646	7,7	226	6,7
Inconnu*	2 278	15,9	1 422	14,4	494	12,8

Tableau 10 (suite)

Caractéristique	Inclusion (N=14 483)		2 ans (âge corrigé) (N=9 958)		5 ans (N=3 898)	
	n	%	n	%	n	%
Nombre de frères et sœurs à la naissance						
0	3 632	31,9	2 467	31,9	992	34,8
1	3 638	31,9	2 494	32,3	910	31,9
2 ou plus	4 119	36,2	2 762	35,8	950	33,3
Inconnu*	2 940	20,5	2 142	21,7	1 014	26,2
Statut marital de la mère						
En couple	11 563	87,3	7 984	87,0	3 191	87,4
Célibataire	1 683	12,7	1 197	13,0	460	12,6
Inconnu*	1 237	8,5	777	7,8	247	6,3
Sexe de l'enfant						
Féminin	6 915	47,7	4 718	47,4	1 809	46,4
Masculin	7 568	52,3	5 240	52,6	2 089	53,6
Age gestationnel (semaines d'aménorrhée)						
<=28	4 669	32,2	3 192	32,1	1 262	32,4
29-30	3 872	26,7	2 684	27,0	1 060	27,2
31-32	5 942	41,0	4 082	41,0	1 576	40,4
Z-score du périmètre crânien à la naissance						
<-1	1 458	17,5	1 020	17,7	409	18,2
[-1;0[2 839	34,1	1 974	34,3	747	33,3
>0	4 022	48,3	2 753	47,9	1 087	48,5
Inconnu*	6 164	42,6	4 211	42,3	1 655	42,5
Z-score du poids à la naissance						
<-1	2 576	17,9	1 785	18,0	734	18,9
[-1;0[5 847	40,7	4 020	40,6	1 568	40,4
>0	5 946	41,4	4 087	41,3	1 582	40,7
Inconnu*	114	0,8	66	0,7	14	0,4
Delta du z-score du poids						
<-1	6 614	55,9	5 878	79,0	2 125	82,1
[-1;0[1 858	15,7	1 206	16,2	380	14,7
>0	3 361	28,4	356	4,8	84	3,2
Inconnu*	2 650	18,3	2 518	25,3	1 309	33,6
Prématurité provoquée						
Non	6 353	50,9	4 437	51,3	1 628	49,3
Oui	6 130	49,1	4 215	48,7	1 671	50,7
Inconnu*	2 000	13,8	1 306	13,1	599	15,4
Procréation médicalement assistée						
Non	9 724	85,9	6 646	85,5	2 537	85,3
Oui	1 602	14,1	1 124	14,5	438	14,7
Inconnu*	3 157	21,8	2 188	22,0	923	23,7
Naissance multiple						
Non	10 382	71,7	7 075	71,1	2 668	68,4
Oui	4 098	28,3	2 880	28,9	1 230	31,6
Inconnu*	3	0,0	3	0,0	0	0,0

Tableau 10 (suite)

Caractéristique	Inclusion (N=14 483)		2 ans (âge corrigé) (N=9 958)		5 ans (N=3 898)	
	n	%	n	%	n	%
Age de la mère à la naissance (années)						
<20	179	1,3	133	1,4	62	1,6
20-29	4 489	31,4	3 153	32,1	1 261	32,9
30-39	7 536	52,7	5 042	51,4	1 912	49,8
>=40	2 085	14,6	1 483	15,1	602	15,7
Inconnu	194	1,3	147	1,5	61	1,6
Mode d'accouchement						
Voie basse	5 141	36,8	3 531	36,8	1 289	34,7
Césarienne	8 812	63,2	6 052	63,2	2 424	65,3
Inconnu*	530	3,7	375	3,8	185	4,7
Tabac durant la grossesse						
Non	11 343	92,2	7 980	91,6	3 201	90,4
Oui	959	7,8	729	8,4	338	9,6
Inconnu*	2 181	15,1	1 249	12,5	359	9,2
Alcool durant la grossesse						
Non	12 090	99,2	8 560	99,2	3 477	98,9
Oui	95	0,8	73	0,8	37	1,1
Inconnu*	2 298	15,9	1 325	13,3	384	9,9
Cannabis durant la grossesse						
Non	12 042	99,2	8 519	99,3	3 447	99,4
Oui	92	0,8	62	0,7	20	0,6
Inconnu*	2 349	16,2	1 377	13,8	431	11,1
Autre produit toxique durant la grossesse						
Non	11 769	99,7	8 305	99,7	3 345	99,5
Oui	32	0,3	24	0,3	16	0,5
Inconnu*	2 682	18,5	1 629	16,4	537	13,8
Produit de substitution durant la grossesse						
Non	11 939	99,8	8 440	99,8	3 407	99,6
Oui	24	0,2	21	0,2	13	0,4
Inconnu*	2 520	17,4	1 497	15,0	478	12,3
Corticostéroïdes anténataux						
Non	1 944	15,1	1 287	14,6	503	15,0
Oui	10 897	84,9	7 519	85,4	2 850	85,0
Inconnu*	1 642	11,3	1 152	11,6	545	14,0
Chorioamniotite						
Non	10 226	87,2	7 081	87,4	2 680	87,0
Oui	1 501	12,8	1 022	12,6	400	13,0
Inconnu*	2 756	19,0	1 855	18,6	818	21,0
Résultats auditifs						
Normal	11 817	88,1	8 063	87,7	3 127	87,3
A contrôler	1 043	7,8	747	8,1	310	8,7
Examen non fait	560	4,2	381	4,1	145	4,0
Inconnu*	1 063	7,3	767	7,7	316	8,1

Tableau 10 (suite)

Caractéristique	Inclusion (N=14 483)		2 ans (âge corrigé) (N=9 958)		5 ans (N=3 898)	
	n	%	n	%	n	%
Rétinopathie stade 2 ou plus						
Non	14 106	97,4	9 723	97,6	3 811	97,8
Oui	377	2,6	235	2,4	87	2,2
Corticostéroïdes postnataux						
Non	3 156	87,1	6 982	87,0	2 617	85,7
Oui	1 486	12,9	1 042	13,0	436	14,3
Inconnu*	2 967	20,5	1 934	19,4	845	21,7
Traitement anti-virus respiratoire syncytial						
Non	6 509	45,0	4 559	45,8	1 764	45,3
Oui	7 955	55,0	5 387	54,2	2 130	54,7
Inconnu*	19	0,1	12	0,1	4	0,1
Dysplasie bronchopulmonaire						
Non	11 650	80,6	8 046	80,9	3 118	80,1
Oui	2 812	19,4	1 899	19,1	774	19,9
Inconnu*	21	0,1	13	0,1	6	0,2
Pathologie fœtale additionnelle						
Non	13 668	94,4	9 372	94,2	3 649	93,6
Oui	806	5,6	579	5,8	249	6,4
Inconnu*	9	0,1	7	0,1	0	0,0
Evaluation neurologique						
Pas de risque	931	6,4	667	6,7	244	6,3
Risques faibles	12 356	85,3	8 439	84,7	3 284	84,2
Risques modérés	943	6,5	673	6,8	302	7,7
Risques élevés	253	1,7	179	1,8	68	1,7
Intervention chirurgicale à la naissance						
Non	10 613	90,0	7 353	90,0	2 895	90,2
Oui	1 175	10,0	818	10,0	313	9,8
Inconnu*	2 695	18,6	1 787	17,9	690	17,7
Allaitement prévu à la sortie						
Oui	5 463	47,1	3 950	48,7	1 623	50,2
Non	6 127	52,9	4 160	51,3	1 610	49,8
Inconnu*	2 893	20,0	1 848	18,6	665	17,1
Durée d'hospitalisation (jours)						
0-30	1 402	10,5	944	10,3	357	10,0
31-50	4 541	33,9	3 078	33,7	1 218	34,1
51-70	3 463	25,8	2 373	26,0	891	25,0
>70	3 994	29,8	2 734	29,9	1 103	30,9
Inconnu*	1 083	7,7	829	8,3	329	8,4

Tableau 10 (suite et fin)

Caractéristique	Inclusion (N=14 483)		2 ans (âge corrigé) (N=9 958)		5 ans (N=3 898)	
	n	%	n	%	n	%
Niveau urbain du lieu de vie (nombre d'habitants)						
Moyen ou faible (<50 000)	3 368	26,8	2 403	27,2	1 004	29,7
Elevé (>50 000)	9 180	73,2	6 443	72,8	2 375	70,3
Inconnu*	1 935	13,4	1 112	11,2	519	13,3
Réseau d'inclusion						
RPSOF – Région Sud-Ouest	3 957	27,3	2 676	26,9	976	25,0
RSPP – Paris	2 758	19,0	1 854	18,6	703	18,0
NEF – Région Nord-Est	3 153	21,8	2 190	22,0	824	21,1
RPVM – Région Sud-Est	1 474	10,2	973	9,8	351	9,0
RPVO – Région Nord-Ouest	1 721	11,9	1 241	12,5	502	12,9
MYP A – Région Ouest	1 420	9,8	1 024	10,3	542	13,9
Année de naissance						
2015	1 326	9,2	1 315	13,2	1 301	33,4
2016	2 036	14,1	1 996	20,0	1 948	50,0
2017	2 168	15,0	2 127	21,4	649	16,6
2018	2 047	14,1	2 002	20,1	0	0,0
2019	2 136	14,7	2 113	21,2	0	0,0
2020	2 017	13,9	405	4,1	0	0,0
2021	2 050	14,2	0	0,0	0	0,0
2022	703	4,9	0	0,0	0	0,0
Médecin pilote						
Hospitalier (public/privé)	8 616	64,6	5 940	64,3	2 337	65,3
Libéral	3 612	27,1	2 470	26,7	943	26,3
Autre	1 115	8,4	831	9,0	301	8,4
Inconnu*	1 140	7,9	717	7,2	317	8,1
Type de maternité						
Niveau 3	11 067	80,4	7 581	80,2	2 977	81,2
Niveau 1 ou 2	2 699	19,6	1 870	19,8	689	18,8
Inconnu*	717	5,0	507	5,1	232	6,0
Enfant né en dehors de son établissement d'inclusion						
Non	8 298	59,5	5 773	60,2	2 376	63,8
Oui	5 649	40,5	3 824	39,8	1 350	36,2
Inconnu*	536	3,7	361	3,6	172	4,4
Distance entre le lieu de domicile et le médecin pilote (km)						
<3	6 972	54,0	4 916	54,4	1 906	54,3
3-10	2 824	21,9	1 937	21,5	776	22,1
>10	3 112	24,1	2 177	24,1	830	23,6
Inconnu*	1 575	10,9	928	9,3	386	9,9

* Les proportions de données inconnues sont calculées par rapport au nombre total de sujets, tandis que les proportions dans chaque catégorie d'une variable donnée sont calculées par rapport au nombre de sujets ayant des données complètes pour cette variable.

3.2. Résultats à 2 ans

Dans les analyses univariées comparant les caractéristiques des enfants qui ont participé ou non à la visite des deux ans (tableau 11), nous avons constaté que les enfants dont les parents avaient un niveau socio-économique plus faible étaient davantage susceptibles d'être perdus de vue. Cela se traduisait par un niveau d'éducation inférieur de la mère, une ressource économique autre que le travail d'un parent, une profession des parents autre qu'une profession cadre ou intellectuelle, y compris le congé parental et le chômage, l'absence de couverture médicale ou une couverture médicale pour les faibles revenus, un logement social ou chez un proche, un nombre plus élevé d'enfants dans le foyer, ou enfin un couple séparé.

Parmi les variables cliniques, un âge gestationnel plus faible, la prescription postnatale de corticostéroïdes, la PMA, la naissance multiple, l'accouchement par césarienne, la rétinopathie du prématuré, la dysplasie broncho-pulmonaire, un risque faible ou modéré dans l'évaluation neurologique, l'intervention chirurgicale à la naissance, l'allaitement maternel à la sortie de l'hôpital, une durée d'hospitalisation plus courte et la prescription d'anticorps monoclonaux anti-VRS étaient associés à une plus grande observance. En revanche, un poids de naissance et un périmètre crânien supérieurs à la moyenne (pour le même sexe et le même âge gestationnel), ainsi que le jeune âge de la mère, le tabagisme ou la consommation de cannabis pendant la grossesse étaient plus probables chez les enfants perdus de vue par la suite.

Parmi les variables contextuelles, les enfants nés dans une maternité de niveau 1 ou 2 (par rapport à une maternité de niveau 3), nés en dehors de la maternité d'inclusion ou inclus dans certains réseaux spécifiques étaient davantage susceptibles d'être perdus de vue, tandis que ceux suivis par un médecin pilote libéral ou non hospitalier étaient moins susceptibles d'être perdus de vue. Un niveau d'urbanité plus élevé et une plus grande distance entre le domicile et le médecin pilote étaient également associés à une plus grande observance.

Des associations univariées similaires ont été trouvées dans les analyses avant IM des données manquantes (tableau S10 en annexe 4).

Dans le modèle multivarié final, les facteurs suivants étaient associés significativement à une plus grande probabilité de ne pas se présenter à la visite des deux ans : niveau d'éducation

secondaire par rapport à un niveau supérieur, chômage ou congé parental pour la mère, plus grand nombre d'enfants à la maison, âge maternel plus jeune, tabagisme pendant la grossesse ou accouchement en dehors de la maternité d'inclusion. Les facteurs associés à une diminution du risque d'être perdu de vue à deux ans étaient un âge gestationnel plus faible, la prescription postnatale d'une corticothérapie, l'allaitement maternel à la sortie de l'hôpital, un médecin pilote libéral ou non hospitalier, et une plus grande distance entre le médecin pilote et le lieu de résidence. Aucune variable autre que la durée d'hospitalisation n'a été exclue a posteriori pour cause de colinéarité.

Tableau 11 : Facteurs associés à l'absence de visite à deux ans dans les régressions logistiques univariées et multivariées après imputation multiple, cohorte HYGIE-SEV, 1er novembre 2015-31 octobre 2022 (N=9 958)

Variable	Univarié			Multivarié	
	OR	IC 95%	p	OR	IC 95%
Niveau d'éducation de la mère			<0,001		
Supérieur	1,00	[Référence]		1,00	[Référence]
Secondaire (Collège/CAP/Lycée/BAC)	1,56	[1,42 ; 1,70]		1,19	[1,07 ; 1,33]
Aucun ou niveau primaire	1,79	[1,48 ; 2,16]		1,21	[0,97 ; 1,50]
Ressource économique du foyer			<0,001		
Travail d'un parent au moins	1,00	[Référence]		1,00	[Référence]
Aides sociales	1,60	[1,40 ; 1,84]		1,02	[0,84 ; 1,25]
Autre source	2,06	[1,48 ; 2,86]		1,36	[0,93 ; 1,97]
Aucune	1,41	[1,10 ; 1,82]		0,93	[0,68 ; 1,28]
Situation socio-professionnelle de la mère			<0,001		
Travail : cadre et profession intellectuelle	1,00	[Référence]		1,00	[Référence]
Travail : intermédiaire	1,18	[0,95 ; 1,46]		0,99	[0,79 ; 1,26]
Travail : ouvrier ou employé	1,31	[1,15 ; 1,51]		1,01	[0,86 ; 1,20]
Congé parental	1,76	[1,54 ; 2,02]		1,29	[1,08 ; 1,53]
Chômage/Sans emploi	1,99	[1,72 ; 2,28]		1,35	[1,13 ; 1,62]
Situation socio-professionnelle du père			<0,001		
Travail : cadre et profession intellectuelle	1,00	[Référence]		1,00	[Référence]
Travail : intermédiaire	1,41	[1,20 ; 1,65]		1,14	[0,96 ; 1,36]
Travail : ouvrier ou employé	1,46	[1,30 ; 1,64]		1,08	[0,93 ; 1,24]
Congé parental	2,06	[1,37 ; 3,12]		1,29	[0,83 ; 2,01]
Chômage/Sans emploi	1,87	[1,63 ; 2,15]		1,10	[0,90 ; 1,34]
Couverture sociale			<0,001		
Sécurité sociale	1,00	[Référence]		1,00	[Référence]
Aucune ou AME ou CMU	1,55	[1,38 ; 1,74]		1,11	[0,96 ; 1,29]
Conditions de logement			0,005		
Adéquates	1,00	[Référence]		1,00	[Référence]
Inadéquates (logement trop petit ou insalubre)	1,17	[0,95 ; 1,43]		0,93	[0,74 ; 1,17]
Chez un proche	1,37	[1,14 ; 1,65]		1,00	[0,81 ; 1,26]
Foyer d'accueil, établissement d'urgence ou autre	1,37	[1,13 ; 1,66]		1,07	[0,85 ; 1,35]
Nombre de frères et sœurs au moment de la naissance			<0,001		
0	1,00	[Référence]		1,00	[Référence]
1	1,15	[1,03 ; 1,29]		1,13	[1,00 ; 1,29]
2 ou plus	1,45	[1,30 ; 1,62]		1,36	[1,20 ; 1,55]
Statut marital de la mère			<0,001		
En couple	1,00	[Référence]		1,00	[Référence]
Célibataire	1,33	[1,18 ; 1,51]		1,02	[0,88 ; 1,19]
Sexe			0,826		
Féminin	1,00	[Référence]		1,00	[Référence]
Masculin	1,01	[0,93 ; 1,09]		0,99	[0,91 ; 1,08]
Age gestationnel (semaines d'aménorrhée)			<0,001		
≤28	0,72	[0,65 ; 0,79]		0,77	[0,67 ; 0,88]
29-30	0,82	[0,74 ; 0,91]		0,82	[0,74 ; 0,92]
31-32	1,00	[Référence]		1,00	[Référence]

Tableau 11 (suite)

Variable	Univarié			Multivarié	
	OR	IC 95%	p	OR	IC 95%
Z-score du périmètre crânien à la naissance			0,039		
<-1	1,00	[Référence]		1,00	[Référence]
[-1 ; 0[1,02	[0,89 ; 1,17]		1,04	[0,88 ; 1,23]
>0	1,14	[1,00 ; 1,29]		1,08	[0,90 ; 1,28]
Z-score du poids à la naissance			0,001		
<-1	1,00	[Référence]		1,00	[Référence]
[-1;0[1,11	[0,99 ; 1,25]		1,06	[0,92 ; 1,23]
>0	1,23	[1,09 ; 1,38]		1,17	[0,99 ; 1,39]
Delta du z-score du poids			0,001		
<-1	1,00	[Référence]		1,09	[0,99 ; 1,20]
[-1 ; 0[1,17	[1,07 ; 1,27]		1,20	[0,95 ; 1,53]
>0	1,23	[0,99 ; 1,54]			
Prématurité provoquée			0,138		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,94	[0,86 ; 1,02]		1,05	[0,93 ; 1,19]
Procréation médicalement assistée			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,68	[0,59 ; 0,78]		0,87	[0,74 ; 1,01]
Naissance multiple			0,008		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,89	[0,80 ; 0,97]		0,98	[0,88 ; 1,09]
Age de la mère à la naissance (années)			<0,001		
<20	2,20	[1,55 ; 3,12]		1,64	[1,12 ; 2,39]
20-29	1,23	[1,12 ; 1,35]		1,20	[1,08 ; 1,33]
30-39	1,00	[Référence]		1,00	[Référence]
>=40	0,96	[0,85 ; 1,08]		0,92	[0,81 ; 1,05]
Mode d'accouchement			<0,001		
Voie basse	1,00	[Référence]		1,00	[Référence]
Césarienne	0,86	[0,79 ; 0,93]		0,89	[0,78 ; 1,01]
Tabac durant la grossesse			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	1,54	[1,32 ; 1,79]		1,35	[1,13 ; 1,61]
Alcool durant la grossesse			0,051		
Non	1,00	[Référence]		1,00	[Référence]
Oui	1,53	[1,00 ; 2,34]		1,07	[0,63 ; 1,80]
Cannabis durant la grossesse			0,018		
Non	1,00	[Référence]			
Oui	1,71	[1,09 ; 2,68]		1,17	[0,67 ; 2,03]
Autre produit toxique durant la grossesse			0,558		
Non	1,00	[Référence]			
Oui	1,10	[0,59 ; 2,03]			
Produit de substitution durant la grossesse			0,488		
Non	1,00	[Référence]			
Oui	1,06	[0,53 ; 2,11]			
Corticostéroïdes anténataux			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,79	[0,70 ; 0,89]		0,90	[0,79 ; 1,03]

Tableau 11 (suite)

Variable	Univarié			Multivarié	
	OR	IC 95%	p	OR	IC 95%
Chorioamniotite			0,633		
Non	1,00	[Référence]			
Oui	1,03	[0,90 ; 1,18]			
Résultats auditifs			0,609		
Normal	1,00	[Référence]			
A contrôler	0,99	[0,85 ; 1,15]			
Examen non fait	1,11	[0,90 ; 1,37]			
Rétinopathie stade 2 ou plus			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,58	[0,43 ; 0,77]		0,88	[0,64 ; 1,20]
Corticostéroïdes postnatale			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,60	[0,52 ; 0,69]		0,78	[0,65 ; 0,93]
Traitement anti-virus respiratoire syncytial			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,86	[0,79 ; 0,93]		0,92	[0,83 ; 1,02]
Dysplasie bronchopulmonaire			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,67	[0,60 ; 0,74]		0,87	[0,72 ; 1,01]
Pathologie fœtale additionnelle			0,483		
Non	1,00	[Référence]			
Oui	1,06	[0,90 ; 1,26]			
Evaluation neurologique			0,042		
Pas de risque	1,00	[Référence]		1,00	[Référence]
Risques faibles	0,82	[0,70 ; 0,96]		0,85	[0,72 ; 1,01]
Risques modérés	0,74	[0,59 ; 0,92]		0,84	[0,67 ; 1,06]
Risques élevés	0,85	[0,60 ; 1,19]		0,90	[0,63 ; 1,28]
Intervention chirurgicale à la naissance			0,018		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,84	[0,72 ; 0,97]		0,98	[0,83 ; 1,16]
Allaitement prévu à la sortie			<0,001		
Oui	1,00	[Référence]		1,00	[Référence]
Non	0,85	[0,78 ; 0,93]		0,89	[0,81 ; 0,99]
Durée d'hospitalisation (jours)			<0,001		
0-30	1,00	[Référence]			
31-50	0,78	[0,68 ; 0,91]			
51-70	0,62	[0,53 ; 0,72]			
> 70	0,53	[0,46 ; 0,61]			
Niveau urbain du lieu de vie (nombre d'habitants)			<0,001		
Moyen ou faible (<50 000)	1,00	[Référence]		1,00	[Référence]
Elevé (>50 000)	0,83	[0,75 ; 0,91]		0,92	[0,81 ; 1,05]
Réseau d'inclusion			<0,001		
RPSOF – Région Sud-Ouest	1,00	[Référence]		1,00	[Référence]
RSPP – Paris	1,44	[1,27 ; 1,63]		1,56	[1,36 ; 1,78]
NEF – Région Nord-Est	1,50	[1,33 ; 1,69]		1,23	[1,08 ; 1,39]
RPVM – Région Sud-Est	1,01	[0,87 ; 1,19]		1,14	[0,96 ; 1,36]
RPVO – Région Nord-Ouest	2,70	[2,35 ; 3,10]		2,54	[2,17 ; 2,98]
MYPA – Région Ouest	1,11	[0,95 ; 1,30]		1,30	[1,08 ; 1,56]

Tableau 11 (suite et fin)

Variable	Univarié			Multivarié	
	OR	IC 95%	p	OR	IC 95%
Année de naissance			<0,001		
2015	1,00	[Référence]		1,00	[Référence]
2016	0,89	[0,77 ; 1,03]		0,87	[0,74 ; 1,01]
2017	0,98	[0,85 ; 1,13]		0,97	[0,82 ; 1,13]
2018	1,20	[1,04 ; 1,39]		1,16	[1,00 ; 1,36]
2019-2020	1,06	[0,93 ; 1,22]		1,03	[0,89 ; 1,20]
Médecin pilote			<0,001		
Hospitalier (public/privé)	1,00	[Référence]		1,00	[Référence]
Libéral	0,65	[0,59 ; 0,72]		0,71	[0,63 ; 0,79]
Autre	0,74	[0,64 ; 0,86]		0,65	[0,54 ; 0,77]
Type de maternité			<0,001		
Niveau 3	1,00	[Référence]		1,00	[Référence]
Niveau 1 ou 2	1,31	[1,18 ; 1,45]		1,11	[0,98 ; 1,25]
Enfant né en dehors de son établissement d'inclusion			0,003		
Non	1,00	[Référence]		1,00	[Référence]
Oui	1,14	[1,05 ; 1,24]		1,19	[1,09 ; 1,31]
Distance entre le lieu de domicile et le médecin pilote (km)			0,006		
<3	1,00	[Référence]		1,00	[Référence]
3-10	0,93	[0,83 ; 1,03]		0,88	[0,79 ; 0,99]
>10	0,85	[0,76 ; 0,94]		0,83	[0,74 ; 0,93]

Abréviations: IC, Intervalle de confiance; OR, Odds Ratio

3.3. Résultats à 5 ans

Lors de la visite des cinq ans, nous avons observé des associations univariées similaires à celles trouvées dans les analyses de la visite des deux ans, après (tableau 12) comme avant IM (tableau S11 en annexe 4). Dans le modèle multivarié final, les facteurs associés à une probabilité accrue de ne pas se présenter à la visite des cinq ans étaient le faible niveau d'éducation de la mère, les aides sociales pour faible revenu, la mère en congé parental, le père exerçant une profession intermédiaire, le nombre plus élevé d'enfants dans le foyer, l'âge maternel plus jeune et le tabagisme pendant la grossesse. Les facteurs associés à une meilleure observance étaient un âge gestationnel plus faible, la prescription postnatale de corticostéroïdes, un médecin pilote autre qu'hospitalier et une plus grande distance entre le domicile et le médecin pilote. Aucune variable autre que la durée d'hospitalisation n'a été exclue a posteriori pour cause de colinéarité.

3.4. Analyse de sensibilité

Dans l'analyse de sensibilité utilisant une définition plus stricte de la perte de suivi à la visite des deux ans (tableau S12 en annexe 4), le fait que le père exerce une profession de cadre ou intellectuelle, que l'enfant soit logé dans un foyer social ou un établissement social d'urgence et que la mère ait fumé pendant la grossesse étaient associés à une plus grande probabilité de perte de suivi, tandis qu'un âge gestationnel plus faible, la PMA, la prescription anténatale de corticostéroïdes et le fait d'avoir un médecin pilote autre qu'hospitalier, étaient associés à une probabilité plus faible.

Tableau 12 : Facteurs associés à l'absence de visite à cinq ans dans les régressions logistiques univariées et multivariées après imputation multiple, cohorte HYGIE-SEV, 1er novembre 2015-31 octobre 2022 (N=3 898)

Variable	Univarié			Multivarié	
	OR	IC 95%	p	OR	IC 95%
Niveau d'éducation de la mère			<0,001		
Supérieur	1,00	[Référence]		1,00	[Référence]
Secondaire (Collège/CAP/Lycée/BAC)	1,62	[1,41 ; 1,87]		1,28	[1,07 ; 1,52]
Aucun ou niveau primaire	1,43	[1,04 ; 1,97]		1,11	[0,76 ; 1,62]
Ressource économique du foyer			<0,001		
Travail d'un parent au moins	1,00	[Référence]		1,00	[Référence]
Aides sociales	1,82	[1,41 ; 2,34]		1,44	[1,01 ; 2,05]
Autre source	1,31	[0,78 ; 2,20]		1,12	[0,62 ; 2,02]
Aucune	1,39	[0,89 ; 2,17]		1,11	[0,62 ; 1,96]
Situation socio-professionnelle de la mère			<0,001		
Travail : cadre et profession intellectuelle	1,00	[Référence]		1,00	[Référence]
Travail : intermédiaire	1,21	[0,88 ; 1,66]		0,89	[0,62 ; 1,25]
Travail : ouvrier ou employé	1,27	[1,04 ; 1,56]		0,95	[0,74 ; 1,21]
Congé parental	1,93	[1,56 ; 2,39]		1,36	[1,04 ; 1,77]
Chômage/Sans emploi	1,73	[1,39 ; 2,16]		1,07	[0,81 ; 1,42]
Situation socio-professionnelle du père			<0,001		
Travail : cadre et profession intellectuelle	1,00	[Référence]		1,00	[Référence]
Travail : intermédiaire	1,69	[1,32 ; 2,18]		1,47	[1,12 ; 1,94]
Travail : ouvrier ou employé	1,52	[1,28 ; 1,81]		1,21	[0,97 ; 1,51]
Congé parental	3,18	[1,45 ; 6,96]		1,85	[0,82 ; 4,20]
Chômage/Sans emploi	1,76	[1,42 ; 2,20]		1,13	[0,83 ; 1,54]
Couverture sociale			0,005		
Sécurité sociale	1,00	[Référence]		1,00	[Référence]
Aucune ou AMÉ ou CMU	1,38	[1,10 ; 1,75]		0,98	[0,75 ; 1,28]
Conditions de logement			0,003		
Adéquates	1,00	[Référence]		1,00	[Référence]
Inadéquates (logement trop petit ou insalubre)	0,94	[0,73 ; 1,20]		0,73	[0,55 ; 0,97]
Chez un proche	1,56	[1,18 ; 2,07]		1,20	[0,86 ; 1,68]
Foyer d'accueil, établissement d'urgence ou autre	1,39	[1,03 ; 1,88]		1,20	[0,84 ; 1,71]
Nombre de frères et sœurs à la naissance			<0,001		
0	1,00	[Référence]		1,00	[Référence]
1	1,05	[0,88 ; 1,25]		1,09	[0,89 ; 1,32]
2 ou plus	1,37	[1,15 ; 1,64]		1,47	[1,18 ; 1,82]
Statut marital de la mère			0,070		
En couple	1,00	[Référence]		1,00	[Référence]
Célibataire	1,20	[0,98 ; 1,47]		0,85	[0,66 ; 1,09]
Sexe de l'enfant			0,574		
Féminin	1,00	[Référence]		1,00	[Référence]
Masculin	1,04	[0,91 ; 1,18]		1,02	[0,89 ; 1,17]

Tableau 12 (suite)

Variable	Univarié			Multivarié	
	OR	IC 95%	p	OR	IC 95%
Age gestationnel (semaines d'aménorrhée)			<0,001		
≤28	0,74	[0,64 ; 0,86]		0,77	[0,62 ; 0,96]
29-30	0,81	[0,69 ; 0,95]		0,85	[0,70 ; 0,99]
31-32	1,00	[Référence]		1,00	[Référence]
Z-score du périmètre crânien à la naissance			0,121		
<-1	1,00	[Référence]		1,00	[Référence]
[-1 ; 0[1,19	[0,96 ; 1,47]		1,23	[0,94 ; 1,61]
>0	1,22	[1,00 ; 1,49]		1,21	[0,92 ; 1,60]
Z-score du poids à la naissance			0,173		
<-1	1,00	[Référence]		1,00	[Référence]
[-1 ; 0[1,14	[0,95 ; 1,36]		1,04	[0,83 ; 1,31]
>0	1,18	[0,99 ; 1,42]		1,07	[0,82 ; 1,39]
Delta du z-score du poids			0,006		
<-1	1,00	[Référence]		1,00	[Référence]
[-1 ; 0[1,12	[0,98 ; 1,36]		1,01	[0,86 ; 1,41]
>0	1,73	[1,21 ; 2,49]		1,71	[1,17 ; 2,51]
Prématurité provoquée			0,180		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,91	[0,79 ; 1,04]		0,90	[0,73 ; 1,11]
Procréation médicalement assistée			0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,81	[0,67 ; 0,99]		1,03	[0,82 ; 1,30]
Naissance multiple			0,896		
Non	1,00	[Référence]			
Oui	0,99	[0,86 ; 1,14]			
Age de la mère à la naissance (années)			<0,001		
<20	3,03	[1,60 ; 5,72]		2,93	[1,47 ; 5,85]
20-29	1,48	[1,27 ; 1,72]		1,49	[1,26 ; 1,77]
30-39	1,00	[Référence]		1,00	[Référence]
≥40	1,14	[0,95 ; 1,38]		1,14	[0,83 ; 1,26]
Mode d'accouchement			0,157		
Voie basse	1,00	[Référence]		1,00	[Référence]
Césarienne	0,90	[0,79 ; 1,04]		1,02	[0,83 ; 1,26]
Tabac durant la grossesse			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	1,55	[1,21 ; 1,98]		1,47	[1,12 ; 1,93]
Alcool durant la grossesse			0,163		
Non	1,00	[Référence]		1,00	[Référence]
Oui	1,53	[0,83 ; 2,80]		1,18	[0,55 ; 2,54]
Cannabis durant la grossesse			0,141		
Non	1,00	[Référence]		1,00	[Référence]
Oui	1,83	[0,79 ; 4,23]		1,18	[0,42 ; 3,31]
Autre produit toxique durant la grossesse			0,306		
Non	1,00	[Référence]			
Oui	1,42	[0,62 ; 3,28]			
Produit de substitution durant la grossesse			0,588		
Non	1,00	[Référence]			
Oui	1,06	[0,50 ; 2,24]			

Tableau 12 (suite)

Variable	Univarié			Multivarié	
	OR	IC 95%	p	OR	IC 95%
Corticostéroïdes anténataux			0,054		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,83	[0,68 ; 1,00]		0,91	[0,73 ; 1,13]
Chorioamniotite			0,252		
Non	1,00	[Référence]			
Oui	0,88	[0,71 ; 1,09]			
Résultats auditifs			0,652		
Normal	1,00	[Référence]			
A contrôler	0,90	[0,71 ; 1,14]			
Examen non fait	0,98	[0,69 ; 1,38]			
Rétinopathie stade 2 ou plus			0,540		
Non	1,00	[Référence]			
Oui	0,87	[0,57 ; 1,35]			
Corticostéroïdes postnataux			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,62	[0,51 ; 0,75]		0,78	[0,61 ; 0,99]
Traitement anti-virus respiratoire syncytial			0,002		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,82	[0,72 ; 0,93]		0,88	[0,75 ; 1,05]
Dysplasie bronchopulmonaire			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,74	[0,63 ; 0,87]		0,98	[0,79 ; 1,22]
Pathologie fœtale additionnelle			0,516		
Non	1,00	[Référence]			
Oui	0,92	[0,71 ; 1,19]			
Evaluation neurologique			0,327		
Pas de risque	1,00	[Référence]			
Risques faibles	0,79	[0,60 ; 1,04]			
Risques modérés	0,80	[0,56 ; 1,13]			
Risques élevés	0,97	[0,55 ; 1,72]			
Intervention chirurgicale à la naissance			0,500		
Non	1,00	[Référence]			
Oui	0,93	[0,73 ; 1,17]			
Allaitement prévu à la sortie			0,746		
Oui	1,00	[Référence]			
Non	0,99	[0,87 ; 1,14]			

Tableau 12 (suite et fin)

Variable	Univarié			Multivarié	
	OR	IC 95%	p	OR	IC 95%
Intervention chirurgicale à la naissance			0,500		
Non	1,00	[Référence]			
Oui	0,93	[0,73 ; 1,17]			
Allaitement prévu à la sortie			0,746		
Oui	1,00	[Référence]			
Non	0,99	[0,87 ; 1,14]			
Durée d'hospitalisation (jours)			<0,001		
0-30	1,00	[Référence]			
31-50	0,76	[0,59 ; 0,98]			
51-70	0,72	[0,55 ; 0,92]			
>70	0,58	[0,45 ; 0,75]			
Niveau urbain du lieu de vie (nombre d'habitants)			0,057		
Moyen ou faible (<50 000)	1,00	[Référence]		1,00	[Référence]
Elevé (>50 000)	0,86	[0,74 ; 1,00]		0,97	[0,77 ; 1,21]
Réseau d'inclusion			<0,001		
RPSOF – Région Sud-Ouest	1,00	[Référence]		1,00	[Référence]
RSPP – Paris	1,23	[1,00 ; 1,50]		1,29	[1,04 ; 1,60]
NEF – Région Nord-Est	1,00	[0,83 ; 1,21]		0,77	[0,63 ; 0,95]
RPVM – Région Sud-Est	1,10	[0,86 ; 1,41]		1,25	[0,94 ; 1,66]
RPVO – Région Nord-Ouest	2,91	[2,26 ; 3,75]		3,04	[2,27 ; 4,06]
MYP A – Région Ouest	0,87	[0,71 ; 1,08]		1,22	[0,92 ; 1,62]
Année de naissance			<0,001		
2015	1,00	[Référence]		1,00	[Référence]
2016-2017	1,33	[1,16 ; 1,52]		1,38	[1,17 ; 1,60]
Médecin pilote			<0,001		
Hospitalier (public/privé)	1,00	[Référence]		1,00	[Référence]
Libéral	0,66	[0,57 ; 0,76]		0,70	[0,58 ; 0,83]
Autre	0,83	[0,65 ; 1,06]		0,62	[0,46 ; 0,82]
Type de maternité			0,021		
Niveau 3	1,00	[Référence]		1,00	[Référence]
Niveau 1 ou 2	1,23	[1,03 ; 1,46]		1,07	[0,87 ; 1,32]
Enfant né en dehors de son établissement d'inclusion			0,498		
Non	1,00	[Référence]			
Oui	1,05	[0,91 ; 1,20]			
Distance entre le lieu de domicile et le médecin pilote (km)			0,092		
<3	1,00	[Référence]		1,00	[Référence]
3-10	0,91	[0,77 ; 1,08]		0,83	[0,69 ; 1,00]
>10	0,83	[0,71 ; 0,98]		0,81	[0,67 ; 0,97]

Abréviations: IC, Intervalle de confiance; OR, Odds Ratio

4. Discussion

4.1. Principaux résultats

En analysant cette cohorte de grande taille d'enfants prématurés, nous avons constaté que la perte de suivi était associée à un large éventail de caractéristiques socio-économiques et cliniques individuelles, ainsi qu'à des facteurs contextuels. Les taux de perdus de vue aux visites des deux et cinq ans étaient généralement plus élevés chez les enfants issus d'un milieu socio-économique défavorisé et chez ceux qui étaient en meilleure santé à la naissance.

4.2. Comparaison avec la littérature

Dans l'ensemble, le taux d'attrition dans la cohorte HYGIE-SEV était de 39,0 % à la visite des deux ans (ou 31,3 % selon une définition plus stricte) et de 62,6 % à la visite des cinq ans. Bien que les populations varient d'une étude à l'autre en ce qui concerne le degré de prématurité (extrême, grande ou prématurité modérée à tardive), ces taux s'avèrent plus élevés que dans d'autres études. Ainsi, une analyse systématique de 20 études menées dans neuf pays a révélé que les taux d'attrition à 18 ou 24 mois variaient entre 0 et 28,4 % chez les enfants extrêmement prématurés (nés avant 28 SA) (74). Une étude portant sur les facteurs de non-observance à deux ans chez les enfants grands prématurés (nés avant 33 SA) a révélé un taux d'attrition de 26,8 % (73), tandis que le taux à cinq ans était de 19,0 % dans la cohorte française EPIPAGE-2 dédiée aux enfants grands prématurés (77). Enfin, un taux d'attrition de 10,5 % à deux ans et de 21,9 % à cinq ans a été rapporté dans une autre cohorte pédiatrique française d'enfants prématurés (nés avant 37 SA) (76). Les taux plus élevés observés dans notre étude peuvent être dus au fait que la cohorte issue du programme HYGIE-SEV est une cohorte opérationnelle, en vie réelle, plutôt qu'une cohorte de recherche traditionnelle qui met l'accent sur le suivi des participants.

La plupart des facteurs que nous avons trouvés associés à l'attrition sont cohérents avec ceux rapportés dans la littérature. Dans une revue systématique récente de 46 cohortes d'enfants prématurés (75), les caractéristiques les plus décrites (dans trois cohortes ou plus) étaient les suivantes : mère plus jeune, née à l'étranger (variable non disponible dans HYGIE-SEV) ou

multipare, statut socio-économique inférieur, faible niveau d'éducation de la mère, parent célibataire, âge gestationnel inférieur, meilleure capacité pulmonaire, performances cognitives inférieures et handicaps sévères. De même, dans notre étude, un environnement socio-économique défavorable se traduisait par des mères ayant un niveau d'éducation plus faible ou ne travaillant pas. Nous avons également constaté que l'âge maternel plus jeune était associé à un risque plus élevé de perte de vue que l'âge supérieur à 30 ans. La non réalisation des visites à deux ans et cinq ans plus élevée dans les foyers avec plus d'enfants reflète probablement une plus grande charge de travail pour le ménage et donc des difficultés à participer régulièrement au programme HYGIE-SEV. Ces résultats concernant les facteurs socio-économiques ont également été observés dans des études similaires (71,73,77,79).

Une meilleure observance chez les enfants nés plus prématurément combinée à un taux d'attrition plus faible chez les enfants ayant reçu une prescription postnatale de corticostéroïdes suggère que les enfants présentant des caractéristiques cliniques plus favorables à la naissance sont davantage susceptibles de manquer leur visite des deux ans. Ces facteurs cliniques associés à une plus faible attrition ont été identifiés dans le passé (66,72,77,79,80,180). Une naissance en dehors de la maternité d'inclusion, qui correspond à un transfert du nouveau-né de l'hôpital d'origine vers une maternité plus adaptée au suivi et aux soins, est un inconvénient supplémentaire pour les parents qui peuvent percevoir négativement l'inclusion dans le programme HYGIE-SEV et donc prêter moins d'attention au suivi du programme. Dans le même ordre d'idées, plusieurs études ont montré que le suivi pouvait être influencé par le lieu de naissance et le niveau de soins de la maternité (66,73,79).

D'un point de vue géographique, les enfants vivant à plus de cinq kilomètres de leur site de suivi avaient une meilleure observance que les enfants moins éloignés. Dans la région Ile-de-France, une plus grande distance pourrait signifier que ces familles peuvent se déplacer avec leur propre véhicule et ne dépendent pas des transports publics, ce qui limite leurs contraintes de déplacement. Comme explication alternative, malgré l'ajustement multivarié, nous ne pouvons pas exclure la possibilité que les enfants vivant loin des sites de suivi soient aussi ceux qui présentent d'autres caractéristiques individuelles ou structurelles associées à un meilleur suivi (par exemple, le réseau RPSOF dans nos données). La différence d'attrition entre les réseaux d'inclusion, le type de médecin pilote et le niveau de soins de la maternité reflète

des disparités territoriales. Même si ces effets sont moins prononcés à la visite des cinq ans qu'à celle des deux ans, et encore moins dans l'analyse de sensibilité avec une définition plus stricte, cela démontre l'importance d'ajuster les modèles explicatifs pour ces facteurs, qui peuvent jouer un rôle crucial dans les études de suivi.

4.3. Forces et limites de l'étude

La taille importante de l'échantillon constitue un atout majeur. Au moment des analyses, plus de 28 000 enfants étaient inscrits au programme HYGIE-SEV. La présente étude a inclus 9 958 et 3 898 enfants lors des visites à deux et cinq ans, respectivement, ce qui constitue l'une des plus grandes cohortes d'enfants grands prématurés. Bien que les résultats obtenus lors de la visite à cinq ans soient généralement comparables à ceux obtenus lors de la visite à deux ans, le manque de puissance statistique dû à la taille réduite de l'échantillon peut avoir empêché certains facteurs d'être statistiquement significatifs lors de la visite à cinq ans. Quoiqu'il en soit, le suivi à long terme des enfants est un point fort de l'étude et nous permettra d'étudier les résultats pertinents, tels que le développement cognitif et linguistique, la scolarisation ou les difficultés psycho-émotionnelles. Même si certaines variables sont manquantes, notamment en ce qui concerne la santé de la mère, l'origine géographique et les pathologies pendant la grossesse, un autre point fort de l'étude est le grand nombre et la variété de variables disponibles. L'absence de groupes de données ou de dimensions spécifiques est souvent citée comme une limite d'autres études évaluant les facteurs associés à un suivi optimal (66,71,72,186).

La nature opérationnelle de la cohorte HYGIE-SEV pourrait entraîner une baisse de la qualité des données collectées, notamment en termes de fréquence des données manquantes. Cependant, en utilisant une méthodologie appropriée qui suppose que les données sont manquantes au hasard, nous avons pu limiter la perte d'observations et le biais de sélection qui pourrait en résulter (187–189).

La sous-déclaration de raisons spécifiques de sortie du réseau, telles que le déménagement en dehors de la région Ile-de-France, pourrait conduire à une légère surestimation des taux d'attrition. Les enfants souffrant de handicaps sévères, tels que l'infirmité motrice cérébrale,

les troubles cognitifs sévères ou les problèmes comportementaux graves, et qui sont suivis dans des programmes plus appropriés que le programme HYGIE-SEV, pourraient également expliquer le taux d'abandon. Cependant, seulement 79 des 14 483 enfants ont été identifiés comme ayant quitté le programme pour cette raison. En l'absence d'information sur le motif d'abandon de leur réseau, d'autres enfants sont considérés comme perdus de vue alors qu'ils sont pris en charge dans un autre réseau. Cette sous-déclaration entraîne inévitablement un biais car ces enfants sont généralement plus fragiles sur le plan clinique (190). Cependant, nous pensons que leur nombre est suffisamment faible pour ne pas fausser les résultats. En effet, ce sont les enfants en meilleure santé, et non ceux présentant des pathologies graves, qui étaient les plus susceptibles de manquer leur visite à deux ou cinq ans.

Une autre source possible de surestimation du taux d'attrition est que les médecins chargés du suivi peuvent avoir choisi de ne pas saisir les détails des visites, en particulier pour les enfants en bonne santé. Cela pourrait expliquer en partie pourquoi une moins bonne observance a été associée à des réseaux spécifiques et à un type de médecin. Cependant, nous pensons que le nombre d'enfants concernés est limité et qu'il ne peut expliquer à lui seul l'association entre une meilleure santé de l'enfant et une meilleure observance.

Enfin, il convient de ne pas sur interpréter les résultats compte tenu de la multiplicité des tests, qui augmente le risque d'un résultat statistiquement significatif par hasard. Néanmoins, les variables que nous avons trouvées significativement associées à la perte de suivi sont cohérentes avec celles rapportées dans la littérature.

4.4. Perspectives

Cette étude met en évidence les raisons possibles de la perte de suivi dans les cohortes d'enfants prématurés. La compréhension des mécanismes sous-jacents qui conduisent à l'attrition aidera les réseaux périnataux et les médecins traitants à faire des efforts ciblés pour maintenir les enfants dans le suivi. En pratique, des efforts devraient être faits pour améliorer le suivi des groupes à haut risque identifiés dans cette étude. Il est important de noter que nos résultats suggèrent que ces mécanismes sont multidimensionnels et doivent être pris en compte simultanément.

Cette étude introduit bien la prochaine étude présentée ci-après sur les prédictions individuelles de la perte de vue. En effet, la compréhension des mécanismes liés à la perte de vue telle qu'illustrée ici permettra de mettre en perspective les résultats des prédictions présentés dans le chapitre qui suit.

CHAPITRE 4 –

OPTIMALITE DU SUIVI : ANALYSES

PREDICTIVES

Ce quatrième chapitre présente les résultats d'un troisième travail de thèse s'intéressant au développement d'une méthodologie appropriée concernant la prédiction du suivi non optimal dans le programme HYGIE-SEV.

1. Introduction

L'utilisation de l'Intelligence Artificielle (IA) par le biais d'algorithmes de prédiction sur des données massives et omiques dans divers domaines de santé est de plus en plus commune, ces algorithmes ayant déjà prouvé leur efficacité (119). Dans les études cliniques et épidémiologiques traditionnelles, un échantillon restreint de sujets représentatifs de la population est utilisé pour extrapoler des conclusions sur des questions de santé publique (191). Grâce à l'essor des données massives et des algorithmes de prédiction, les capacités techniques et technologiques ont dû et ont su s'adapter pour supporter tous les Systèmes d'Information (SI) liés aux données de santé (192–195). Désormais, s'appuyer sur des données réelles provenant de SI ou de dossiers médicaux électroniques améliore la qualité des résultats obtenus (119,121,122,196). On retrouve des applications, entre autres, en imagerie (197), en oncologie (127,198) ou encore en génomique (199). On trouve aussi des applications de ces modèles pour prédire l'apparition de certaines maladies, telles que la maladie d'Alzheimer (126), la rétinopathie diabétique (124) ou encore la réhospitalisation (200).

Bien que l'IA et les algorithmes de prédiction suscitent également beaucoup d'intérêt en santé périnatale et pédiatrique, leur utilisation est encore limitée. Plusieurs études ont souligné le manque d'informations disponibles dans la littérature (201,202) ou que les travaux initiés

méritent d'être poursuivis pour améliorer les résultats obtenus (203,204). Cependant une récente étude, la seule à ce jour à notre connaissance, s'est intéressée à la prédiction de la perte de vue dans une cohorte d'enfants grands prématurés (205). Les auteurs ont, entre autres, montré que leur modèle de *Random Forests* était celui qui présentait les meilleurs résultats avec une aire sous la courbe *precision-recall* (Sensibilité en fonction de la VPP) de 89,1 % pour la prédiction de l'attrition à deux ans. Les auteurs soulignent la pertinence de travailler dans une cohorte de vie réelle pour étudier l'attrition dans les cohortes de grands prématurés. Tester et comparer plusieurs algorithmes de prédiction afin d'adapter au mieux la qualité des prédictions individuelles est essentiel pour améliorer le suivi. Une validation externe, sur d'autres cohortes d'enfants grands prématurés, serait également une grande plus-value pour la recherche scientifique et confirmer les résultats prometteurs obtenus.

L'objectif de la présente étude était d'explorer différentes techniques de prédiction en s'appuyant sur plusieurs algorithmes de prédiction pour prédire la perte de vue des enfants vulnérables à deux ans et à cinq ans. Dans les parties précédentes nous avons pu comprendre les facteurs qui expliquent ou qui jouent un rôle dans la perte de vue, le but étant d'utiliser ces facteurs à des fins de prédiction. En s'appuyant sur les données disponibles dans HYGIE-SEV, cette nouvelle étude établit une méthodologie qui permet de comparer plusieurs algorithmes de prédiction afin d'en évaluer la capacité prédictive et de sélectionner celui donnant les meilleurs résultats. Le deuxième objectif de cette approche innovante était de fournir aux professionnels de santé qui suivent les enfants dans le programme HYGIE-SEV un outil basé sur les résultats obtenus. Cet outil permettra de prévenir la perte de suivi en tenant compte des caractéristiques socio-économiques, néonatales, liées à la grossesse et contextuelles, contribuant ainsi à l'amélioration du suivi individualisé, en lien avec la médecine personnalisée. Les facteurs les plus prédictifs ont également été décrits afin de déterminer quels sont les paramètres les plus influents sur la perte de vue pour que les médecins pilotes et les investigateurs du programme HYGIE-SEV soient plus attentifs à ces facteurs.

2. Méthodologie

2.1. Population d'étude

Pour cette étude, la population analysée comprenait les grands prématurés (enfants nés avant 33 SA) inclus entre le 1er novembre 2015 et le 31 décembre 2022. Concernant l'analyse à 2 ans, seuls les enfants âgés de plus de deux ans et demi en âge corrigé au 31 décembre 2022 étaient éligibles. De même, pour les analyses à cinq ans, seuls les enfants de plus de cinq ans et demi en âge réel à la même date étaient éligibles. Les enfants décédés, suivis dans un cadre spécialisé ou ayant quitté l'Île-de-France lors d'un déménagement de la famille avant la visite correspondante étaient exclus de l'étude (figure 17).

Les modèles ont été entraînés sur un jeu de données correspondant à 80 % du jeu de données total et les résultats ont été évalués sur les 20 % restants du jeu de données (considérés comme le jeu de données de test). Sur chaque modèle (à deux ans et à cinq ans), afin de garantir une bonne représentativité des proportions, les données ont été stratifiées sur la variable de sortie Y correspondante. La proportion d'enfants n'ayant pas effectué la visite était par conséquent la même dans le jeu d'entraînement et dans le jeu de test.

2.2. Définition de la variable de sortie

Le critère de jugement ou variable de sortie pour cette recherche était la perte de vue ou non-participation d'un enfant à une visite. À deux ans, la perte de vue était définie comme le fait de ne pas s'être rendu à la visite, quel qu'en soit le motif. De même, à cinq ans, la perte de vue était définie comme le fait de ne pas être venu à la visite des cinq ans. La mesure de ce critère permet d'évaluer la capacité des différents modèles de prédiction à identifier les enfants à risque de ne pas bénéficier d'un suivi adéquat.

2.3. Variables prédictives

Un ensemble de 82 variables prédictives a été utilisé dans l'analyse principale. Ces variables incluaient plusieurs dimensions telles que les caractéristiques socio-économiques (17

variables) et démographiques (6 variables), les aspects cliniques néonataux (32 variables) et liés à la grossesse (11 variables), ainsi que les caractéristiques structurelles du réseau HYGIE-SEV (16 variables). Les variables incluses étaient les mêmes pour les analyses à deux ans et à cinq ans. La variable concernant la date de naissance de l'enfant a été transformée de sorte à créer quatre variables : l'année de naissance (considérée comme une variable discrète de 2015 à 2021), le mois de naissance (variable discrète de 1 à 12) en complément de la semaine de naissance (variable continue de 1 à 53) et le jour de naissance dans la semaine (variable discrète de 1 à 7). L'inclusion de ces quatre variables cherche à déterminer quelle échelle de temps serait la plus prédictive, en explorant des échelles plus larges, telles que le mois de l'année, ou des échelles plus fines, comme la semaine de l'année. Au total ce sont 60 variables catégorielles et 25 variables continues.

Il est à noter que l'intégration de variables de suivi améliore les performances des modèles (205). L'objectif initial étant de prédire la perte de vue au moment de l'inclusion dans le programme HYGIE-SEV, les variables de suivi n'ont pas été incluses dans les données d'entrée du modèle. Toutefois, dans les futurs outils d'aide à la décision destinés aux professionnels de santé, il est important de souligner que chaque prédiction peut être effectuée et ajustée pour prédire la prochaine visite. Ainsi, les caractéristiques des visites précédentes pourront être utilisées pour prédire la visite suivante. Dans des analyses secondaires que nous appellerons « analyses incrémentales » et suivant la même méthodologie, nous avons donc réalisé les mêmes modèles prédictifs de la perte de vue à deux ans et à cinq ans en prenant en compte si l'enfant a effectué ou non ses visites précédentes. Ainsi, à deux ans, deux nouvelles variables ont été ajoutées au modèle : « Visite réalisée à 4 mois » et « Visite réalisée à 1 an ». De même à cinq ans ces deux variables ainsi que celles des visites réalisées à 2 ans, 3 ans et 4 ans ont été ajoutées.

2.4. Modèles prédictifs

Pour assurer la reproductibilité et la standardisation du processus de prédiction, le package *Tidymodels* (206) sous R 4.1.1 a été utilisé. *Tidymodels* est une collection de packages spécialisés dans la modélisation et les algorithmes de prédiction. Un large éventail de modèles est disponible sous *Tidymodels*. De plus, ce package peut être interfacé avec d'autres

packages qui facilitent le pré-traitement des données, l'optimisation des hyperparamètres ou encore l'implémentation de validation croisée. Cette compatibilité simplifie et fluidifie le processus de développement et d'écriture des modèles, permettant ainsi de gagner du temps et d'assurer une plus grande cohérence dans l'ensemble du processus. Grâce à *Tidymodels*, la méthodologie employée peut être facilement adaptée et appliquée à d'autres contextes et études similaires, favorisant ainsi la généralisation de cette approche. De plus, l'utilisation de *Tidymodels* garantit la transparence et la clarté du processus d'analyse, ce qui facilite la compréhension et l'interprétation des résultats obtenus.

Le programme informatique d'analyse statistique ayant permis la réalisation de cette étude a été établi dans une logique d'optimisation des ressources. L'intégralité du travail, de la récupération des données brutes jusqu'à la mise en production des modèles finaux en passant par le *data management* et l'élaboration des modèles a été réalisée sur un serveur de calcul RStudio. La manière de programmer a été optimisée de telle sorte à exploiter au maximum les capacités du serveur en utilisant entre autres le calcul en parallèle avec le package *doParallel* de R. En travaillant sur un outil de versionnage (*versionning*) de code (ici GitLab) et en fonctionnalisant les traitements utilisés, le programme final permet de s'adapter à différents changements de paramètres (jeu de données d'entrée, *outcome*, proportion du jeu d'entraînement, nombre de sous-ensembles dans la validation croisée, liste des modèles à utiliser...). Ce travail a permis de tester plusieurs méthodes de traitement et d'adapter facilement chaque choix méthodologique. A l'avenir les programmes générés pourront être réutilisés pour d'autres problématiques sur HYGIE-SEV ou d'autres thématiques de prédiction.

Pour comparer les performances de différents modèles, plusieurs approches ont été explorées, couvrant les modèles paramétriques, semi-paramétriques et non paramétriques. Dans la catégorie des modèles paramétriques, la régression logistique a été utilisée. Les modèles semi-paramétriques comprenaient une régression logistique pénalisée avec la méthode LASSO et un modèle additif généralisé utilisant la méthode *Multivariate Adaptive Regression Splines* (MARS). Quatre modèles non paramétriques ont été employés : les *Random Forests*, les réseaux de neurones, les *Support Vector Machine* (SVM) et une méthode d'*eXtreme Gradient Boosting* (XGBoost). En somme, sept modèles ont été comparés afin de déterminer le plus performant.

Une méthode d'ensemble, combinant tous ces modèles, a également été testée pour évaluer si les résultats pouvaient être améliorés grâce à cette approche. Une méthode LASSO a été utilisée pour sélectionner les modèles les plus performants dans la méthode d'ensemble.

2.5. Pré-traitement des données

Le pré-traitement des données constitue une étape cruciale, tant pour les questions de recherche que pour les considérations industrielles liées à cette thèse. D'une part, cette étape permet d'établir un processus de pré-traitement des données adapté et en adéquation avec la problématique de recherche, dans le but d'optimiser les résultats obtenus. D'autre part, il est important de trouver un équilibre avec les contraintes industrielles, en généralisant et en standardisant le processus, en réduisant au maximum les temps de calcul et en exploitant de manière optimale les ressources matérielles disponibles.

Sous *Tidymodels* une recette (*recipe* en anglais) a été créée à partir des données d'entraînement en spécifiant la variable dépendante Y en fonction de toutes les autres variables. La variable date de naissance était traitée pour extraire les caractéristiques de l'année, du mois, de la semaine, et du jour de naissance. La variable d'année était ensuite convertie en une variable catégorielle. Les variables étaient modifiées pour que les modalités dont la fréquence était inférieure à 1 % soient regroupées en une seule modalité.

Ensuite, l'imputation des données manquantes pour toutes les variables prédictives était effectuée à l'aide de l'algorithme des k plus proches voisins (k-NN) qui a déjà montré son efficacité comparé à d'autres méthodes d'imputation simple (207–209). Pour éviter, d'une part, la complexité du calcul et, d'autre part, les biais potentiels générés par l'imputation k-NN, les variables ayant un taux de valeurs manquantes important ont été exclues automatiquement des analyses. Plusieurs seuils d'exclusion ont été testés (10 %, 20 %, 30 %, 40 % et 50 %). Le seuil d'exclusion qui donnait les meilleurs résultats sur le critère de l'AUC était le seuil de 20 %. Ainsi les variables avec plus de 20 % de valeurs manquantes n'ont pas été incluses dans les analyses (l'exclusion ayant lieu avant les étapes de pré-traitement). Pour déterminer la valeur k , c'est-à-dire le nombre d'observations les plus similaires à l'individu à

imputer, plusieurs valeurs ont été testées : 3, 5 et 10. Toutes ces valeurs donnaient des résultats proches. La valeur $k = 3$ a été retenue.

Les variables à variance nulle ou proche de zéro ont été identifiées et supprimées de la recette. La corrélation maximale entre chaque variable a été fixée à 80 %. Ainsi, et automatiquement le processus de pré-traitement a supprimé le nombre minimal de variables pour que chaque corrélation soit inférieure à ce seuil de 80%.

La transformation Box-Cox (154) a été appliquée à toutes les variables prédictives continues pour améliorer leur normalité. Les variables catégorielles ont été transformées en variables indicatrices (méthode du *one-hot encoding*). Toutes les variables prédictives continues étaient centrées et réduites. Enfin, la méthode *ADaptive SYNthetic* (ADASYN) (210) a été utilisée pour corriger le déséquilibre des classes de la variable dépendante Y . Le principe général est de créer de nouvelles observations synthétiques de la classe minoritaire pour la variable de sortie Y . Une variation aléatoire est ensuite appliquée aux valeurs des variables X . Cette variation est proportionnelle à la différence entre l'observation synthétique générée et les plus proches voisins correspondant aux observations de la classe majoritaire Y .

Le prétraitement est ensuite appliqué successivement sur les données d'entraînement et les données de test.

2.6. Grilles de recherche

Dans le cadre de cette étude, une méthode d'optimisation des hyperparamètres des modèles a été mise en œuvre en utilisant un plan d'expérience basé sur l'hypercube latin (211). L'hypercube latin est une technique de conception expérimentale qui divise l'espace des hyperparamètres en strates, une strate étant un sous-intervalle de l'espace total (déterminé par des bornes) des hyperparamètres. Le nombre de strates correspond à la taille de la grille de recherche. Ce découpage en strates est réalisé de manière uniforme et un seul hyperparamètre est sélectionné (de manière aléatoire) dans chaque strate. Cela permet ainsi de balayer l'ensemble de l'espace en garantissant une répartition des hyperparamètres plus équilibrée au sein de l'espace défini qu'une répartition complètement aléatoire sans strates.

Chaque combinaison d'hyperparamètres a ensuite été évaluée sur les données d'entraînement en utilisant la validation croisée stratifiée sur la variable d'exposition. Dans cette étude, la taille de la grille de recherche a été fixée à 50. Ainsi 50 combinaisons d'hyperparamètres ont été testées pour chaque modèle. D'autres tailles de grille de recherche ont été testées (100, 200, 500) mais celles-ci n'amélioraient pas sensiblement les résultats. La taille de 50 combinaisons était donc satisfaisante pour cette étude et permettait de réduire les temps de calcul par rapport à des tailles plus élevées. Les hyperparamètres qui ont été optimisés sont détaillés dans la partie des résultats.

2.7. Validation croisée

Les hyperparamètres des différents modèles de prédiction ont été optimisés par validation croisée en 5 blocs (*folds* ou sous-ensembles). D'autres nombres de sous-ensembles ont été également testés (10 et 20) mais celui de 5 donnait les meilleurs résultats en se basant sur le critère de l'AUC.

L'utilisation de la validation croisée stratifiée permet de garantir que l'équilibre des classes de l'*outcome* soit préservé dans chaque sous-ensemble, réduisant ainsi les biais potentiels et améliorant l'estimation des performances du modèle (212). Les résultats de la validation croisée sont ensuite utilisés pour identifier les meilleurs hyperparamètres pour chaque modèle, en se basant sur les métriques d'évaluation décrites en section 2.9.

Le choix des hyperparamètres optimaux pour chaque modèle a été basé sur la performance moyenne obtenue au cours de la validation croisée, en utilisant l'AUC. Une fois les hyperparamètres optimaux sélectionnés, les modèles ont été entraînés sur l'ensemble des données d'entraînement en utilisant ces hyperparamètres et évalués sur les données de test pour estimer leurs performances en généralisation. Cette approche d'optimisation des hyperparamètres basée sur l'hypercube latin et la validation croisée stratifiée a permis d'obtenir des modèles performants et robustes, en évitant le sur ajustement et en garantissant une meilleure représentativité des résultats sur l'ensemble des données.

2.8. Evaluation

L'AUC a été définie comme critère de jugement pour évaluer les performances de la validation croisée des modèles d'entraînement. En effet, la courbe ROC représente le taux de vrais positifs en fonction du taux de faux positifs pour différents seuils de classification, permettant ainsi une évaluation globale de la performance du modèle, quelle que soit la répartition des classes dans les données (213).

De plus, l'AUC a une interprétation probabiliste intuitive : elle représente la probabilité qu'un classificateur attribue un score plus élevé à un exemple positif choisi au hasard qu'à un exemple négatif choisi au hasard. Cette caractéristique facilite la comparaison des performances entre différents modèles, car une AUC plus élevée indique une meilleure capacité de discrimination. L'AUC est une mesure indépendante du seuil, ce qui signifie qu'elle prend en compte la performance du modèle à différents seuils de classification. Cela permet d'évaluer la robustesse du modèle sans avoir à fixer un seuil spécifique pour la classification. Par conséquent, l'AUC est considérée comme une mesure de performance plus complète et adaptée à divers scénarios, ce qui en fait notre choix pour évaluer les résultats de la validation croisée.

Pour évaluer les modèles sur les données de test, nous avons également utilisé l'AUC comme critère principal. Cette mesure permet de garantir la comparabilité des résultats entre les différentes étapes de l'évaluation et d'obtenir une estimation fiable de la capacité de généralisation des modèles sélectionnés.

Enfin, d'autres critères d'évaluation ont également été calculés pour évaluer les performances du modèle sur les données de test. Ainsi, la sensibilité, la spécificité, la VPP et la VPN ont été compilées. A noter que ces critères ont été définis pour un seuil de classification à 50 %, c'est-à-dire les enfants dont la probabilité prédite était strictement inférieure à 0,5 ont été classés comme participants à la visite et les enfants dont la probabilité prédite était supérieure ou égale à 0,5 comme non-participants.

2.9. Mesure d'importance des variables prédictives

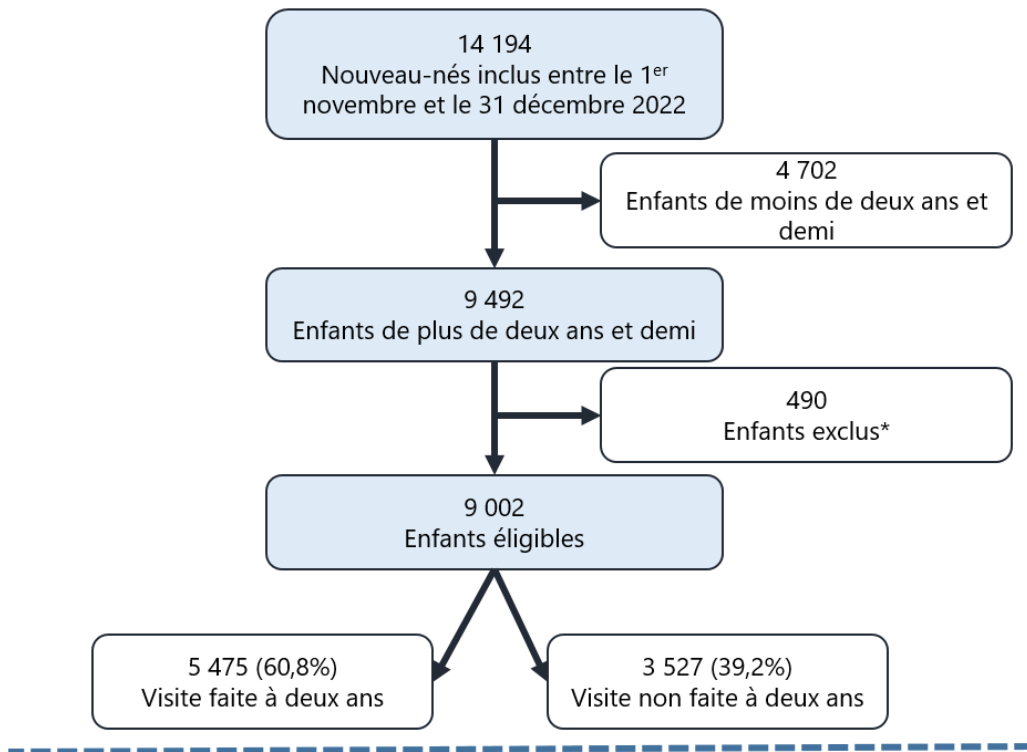
Afin d'explorer des pistes d'interprétation cliniques, l'importance des variables prédictives a été évaluée en utilisant le critère de Gini, tel que formulé par Breiman lors de l'élaboration des arbres de classification (138). Ce chapitre met en évidence les variables qui se sont avérées être les plus prédictives. Bien que l'analyse statistiques prédictive n'offre pas d'explications détaillées sur le rôle spécifique de chaque variable prédictive et la direction de leur association, elle permet néanmoins d'identifier les facteurs qui contribuent le plus à la prédiction. Ces facteurs, bien que devant être interprétés avec prudence, pourraient servir de point de départ pour des recherches approfondies concernant les enfants vulnérables dans le contexte de la médecine personnalisée. Dans nos analyses nous avons mesurer cette valeur d'importance de variables sur le modèle des *Random Forests*.

3. Résultats

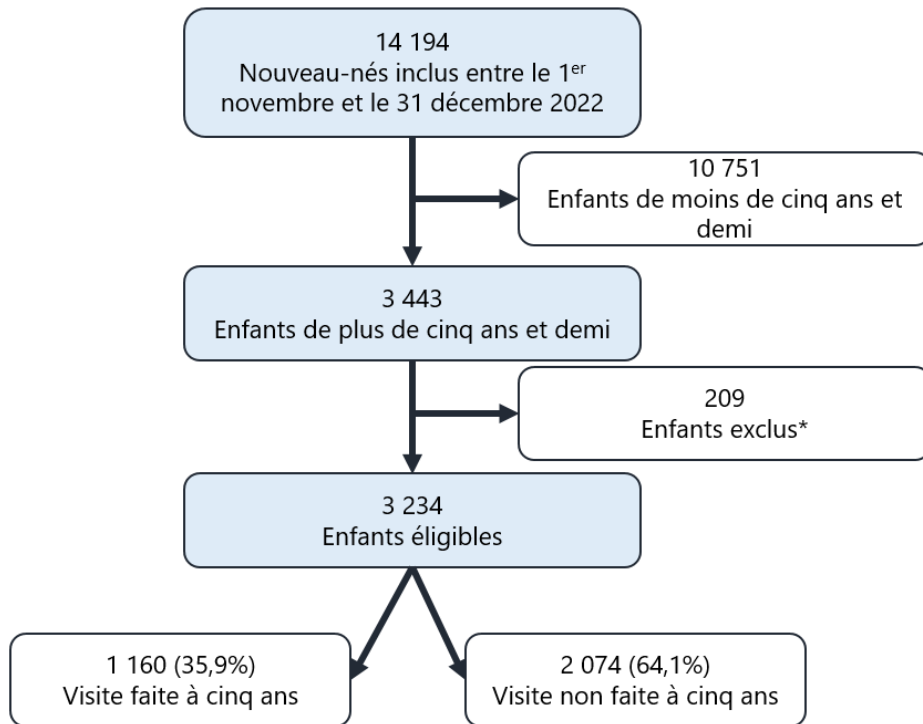
3.1. Sélection de la population

En suivant la méthodologie de sélection de population présentée précédemment, 9 002 enfants étaient inclus dans les analyses à deux ans dont 3 527 (39,2%) enfants n'ayant pas effectué la visite des deux ans. Sur le jeu d'entraînement cela correspondait à 7 201 (80,0 %) enfants dont 2 821 (39,2 %) enfants n'ayant pas réalisé la visite des deux ans et 4 380 (60,8 %) enfants ayant réalisé cette visite. De même sur le jeu de test, 1 801 (20,0 %) enfants ont été inclus avec 706 (39,2 %) enfants n'ayant pas réalisé la visite des deux ans et 1 095 (60,8 %) enfants l'ayant réalisée. A cinq ans, 3 234 enfants étaient éligibles et 2 074 (64,1 %) enfants n'avaient pas réalisé la visite des cinq ans. Sur les données d'entraînement on retrouvait 2 586 (80,0%) enfants dont 1 658 (64,1 %) enfants n'ayant pas réalisé la visite des cinq ans contre 928 (35,9 %) enfants l'ayant réalisée. Sur les données de test, ce sont 648 (20,0 %) qui ont été inclus dans les analyses avec 415 (64,0 %) enfants n'ayant pas réalisée la visite des cinq ans et 233 (36,0 %) ayant réalisé cette visite (figure 17).

Sélection à deux ans



Sélection à cinq ans



*Enfants en suivi spécialisé, décédés ou ayant quitté l'Île-de-France

Figure 17 : Sélection de la population éligible aux analyses prédictives de l'optimalité du suivi

3.2. Résultats à deux ans

3.2.1. Optimisation des hyperparamètres

Grâce à la méthodologie d'optimisation des modèles et via la grille de recherche, les hyperparamètres qui maximisent les performances sur les données d'entraînement ont pu être identifiés. Les hyperparamètres sélectionnés pour chacun des modèles prédictifs de l'absence de visite à deux ans sont présentés dans le tableau 13. Les valeurs fixes des autres hyperparamètres qui n'ont pas été optimisés dans le modèle sont également indiqués.

Tableau 13 : Résultats de l'optimisation des hyperparamètres dans les modèles prédictifs de l'absence de visite à deux ans, cohorte HYGIE-SEV, échantillon d'entraînement (N = 7 201)

Modèle	Hyperparamètre	Valeur	Optimisé
Régression Lasso	penalty	0,004	Oui
MARS	degree	3	Oui
MARS	num_terms	40	Oui
Random Forests	ntrees	2 011	Oui
Random Forests	mtry	9	Oui
Random Forests	min_n	2	Oui
XGBoost	learn_rate	0,3	Non
XGBoost	tree_depth	6	Non
XGBoost	ntrees	1 052	Oui
XGBoost	mtry	9	Oui
XGBoost	min_n	4	Oui
SVM	cost	26,5	Oui
SVM	rbf_sigma	0,01	Oui
Réseaux de neurones	hidden_layers	1	Non
Réseaux de neurones	hidden_units	4	Oui
Réseaux de neurones	activation	softmax	Non
Réseaux de neurones	epochs	55	Oui
Réseaux de neurones	dropout	0,5	Oui

3.2.2. Optimisation de la méthode d'ensemble

Après l'intégration des résultats précédents à la méthode d'ensemble, et grâce à la méthode LASSO, trois modèles ont été sélectionnés (tableau 14) dont les *Random Forests* et le SVM qui

représentaient les poids les plus importants de la méthode d'ensemble. La moyenne de l'AUC était égale à 88,9 % sur les données d'entraînement.

Tableau 14 : Résultats de la méthode d'ensemble pour la prédiction de l'absence de visite à deux ans, cohorte HYGIE-SEV, échantillon d'entraînement (N = 7 201)

Modèle	Poids des modèles
MARS	0,209
SVM	3,660
Random Forests	5,280

3.2.3. Evaluation sur les données de test

Après l'ajustement (*fit*) effectué pour chaque modèle sur l'ensemble du jeu de données d'entraînement, les modèles ont été testés sur les 20 % de données de test. Les différentes métriques évaluées présentent une variation selon le modèle qui est utilisé (tableau 15). En se basant sur le critère de l'AUC, ce sont les *Random Forests* qui donnaient les meilleurs résultats avec une AUC à 77,8 %. Le SVM donnait les deuxièmes meilleurs résultats avec une AUC de 74,9 % suivi du XGBoost avec une AUC de 71,0 %. Les autres modèles donnaient tous une AUC inférieure à 70 %. Enfin la méthode d'ensemble n'améliorait pas les résultats.

Tableau 15 : Evaluation de la performance des modèles prédictifs de l'absence de visite à deux ans, cohorte HYGIE-SEV, échantillon test (N = 1 801)

Modèle	AUC	Sensibilité	Spécificité	VPP	VPN
Random Forests	77,8%	49,8%	86,8%	70,8%	72,9%
SVM	74,9%	44,7%	85,3%	66,2%	70,6%
XGBoost	71,0%	55,3%	76,6%	60,4%	72,7%
Régression logistique pénalisée	68,1%	62,0%	63,4%	52,1%	72,2%
Régression logistique	66,4%	60,0%	61,7%	50,2%	70,6%
MARS	65,6%	36,3%	82,1%	56,4%	66,7%
Réseaux de neurones	63,5%	51,1%	68,3%	50,9%	68,5%
Méthode d'ensemble	77,5%	49,1%	85,9%	69,1%	72,4%

3.2.4. Importance des variables prédictives

Les variables avec la plus grande importance dans le modèle des *Random Forests* étaient entre autres les variables quantitatives (tableau 16) et particulièrement les données cliniques néonatale (durée d'hospitalisation primaire, poids de naissance, poids de sortie, terme de naissance). Mais on retrouve également, en deuxième variable la plus prédictive, la distance entre le lieu de suivi (médecin pilote) et le lieu du domicile (à la naissance). L'âge de la mère au moment de la naissance est la cinquième variable avec la plus haute mesure d'importance. La semaine de naissance du nouveau-né est la septième variable prédictive la plus importante, ce qui souligne l'importance de l'aspect temporel dans la prédiction des perdus de vue. Parmi les premières variables qualitatives, celles d'ordre socio-économique sont les plus importantes avec le niveau d'éducation de la mère, la situation professionnelle de la mère, ou encore le mode de garde à la sortie de l'hôpital. En suivant, le taux de pauvreté puis le niveau de vie de la localité de la famille de l'enfant se situent en 15^{ème} et 16^{ème} position des variables les plus prédictives en y ajoutant une dimension plus socio-démographique que les variables précédentes. Enfin en s'intéressant aux autres variables parmi les 20 variables les plus prédictives, on retrouve d'autres variables socio-économiques comme la profession du père mais aussi des variables cliniques liées à l'accouchement (prématurité induite) ou néonatale (allaitement prévu à l'issue de l'hospitalisation néonatale). Enfin on retrouve également parmi ces variables, le type de médecin suivant l'enfant, ce qui correspond à une dimension plus organisationnelle du programme HYGIE-SEV.

Tableau 16 : Extraction des 20 variables avec la plus grande mesure d'importance dans le modèle des *Random Forests* pour la prédiction de l'absence de visite à deux ans, cohorte HYGIE-SEV, échantillon test (N = 1 801)

Variable	Valeur de la mesure d'importance
Durée d'hospitalisation néonatale	169
Distance entre le domicile et le lieu de suivi	167
Poids de naissance	162
Poids de sortie	155
Age de la mère à la naissance	148
Terme de naissance	148
Semaine de naissance	146
Z score du poids de naissance	100
Niveau d'étude (supérieur)	74
Situation professionnelle de la mère (En activité)	67
Mode de garde envisagé (père ou mère au foyer)	62
Type de médecin (Hospitalier)	58
Profession du conjoint (Employé)	56
Allaitement (Oui)	54
Niveau de vie dans la commune de résidence	54
Taux de pauvreté dans la commune de résidence	54
Type de médecin (Libéral)	52
Sexe de l'enfant (Masculin)	48
Profession du conjoint (Cadre ou profession intellectuelle supérieure)	46
Prématurité provoquée	45

3.3. Résultats à cinq ans

3.3.1. Optimisation des hyperparamètres

Les résultats de l'optimisation des hyperparamètres dans les modèles prédictifs de l'absence de visite à cinq ans (tableau 17) étaient globalement similaires à ceux des modèles à deux ans. On peut noter cependant que, pour le modèle MARS, des fonctions *splines* quadratiques étaient sélectionnées, au lieu de fonctions cubiques à deux ans. Parmi les autres différences on constate également, dans le modèle XGBoost, que les paramètres du nombre de variables sélectionnées aléatoirement à chaque nœud et du nombre d'observations minimales devant être présentes pour qu'un nœud soit divisé (respectivement *mtry* et *min_n*) étaient légèrement inférieurs (respectivement 9 et 4 à deux ans et 4 et 2 à cinq ans), simplifiant ainsi le modèle. Le modèle SVM à cinq ans a été optimisé avec une pénalité associée aux erreurs

de classement moins importante (paramètre *cost* égal à 6,6 contre 26,5 à deux ans). Enfin, pour la méthode des réseaux de neurones, le nombre de fois où les données sont entraînées par le réseau (paramètre *epochs*) a été optimisé à 194 à cinq ans contre 55 à deux ans.

Tableau 17 : Résultats de l'optimisation des hyperparamètres dans les modèles prédictifs de l'absence de visite à cinq ans, cohorte HYGIE-SEV, échantillon d'entraînement (N = 2 586)

Modèle	Hyperparamètre	Valeur	Optimisé
Régression Lasso	penalty	0,006	Oui
MARS	degree	2	Oui
MARS	num_terms	56	Oui
Random Forests	ntrees	2 019	Oui
Random Forests	mtry	9	Oui
Random Forests	min_n	2	Oui
XGBoost	learn_rate	0,3	Non
XGBoost	tree_depth	6	Non
XGBoost	ntrees	1 904	Oui
XGBoost	mtry	4	Oui
XGBoost	min_n	2	Oui
SVM	cost	6,6	Oui
SVM	rbf_sigma	0,03	Oui
Réseaux de neurones	hidden_layers	1	Non
Réseaux de neurones	hidden_units	4	Oui
Réseaux de neurones	activation	softmax	Oui
Réseaux de neurones	epochs	194	Oui
Réseaux de neurones	dropout	0,5	Oui

3.3.2. Optimisation de la méthode d'ensemble

Pour la prédiction de l'absence de visite à cinq ans, seulement deux modèles ont été retenus pour la méthode d'ensemble, les *Random Forests* et le SVM (tableau 18). La moyenne de l'AUC était égale à 93,0 % sur les données d'entraînement.

Tableau 18 : Résultats de la méthode d'ensemble pour la prédiction de l'absence de visite à cinq ans, cohorte HYGIE-SEV, échantillon d'entraînement (N = 2 586)

Modèle	Coefficient stack
SVM	4,030
Random Forests	5,210

3.3.3. Evaluation sur les données de test

Les résultats des modèles prédictifs de l'absence de visite à cinq ans présentaient de moins bonnes performances qu'à deux ans pour les modèles de *Random Forests* mais de meilleures performances pour SVM et XGBoost (tableau 19). Une meilleure AUC pour la méthode d'ensemble a été observée.

Tableau 19 : Evaluation de la performance des modèles prédictifs de l'absence de visite à cinq ans à cinq ans, cohorte HYGIE-SEV, échantillon test (N = 648)

Modèle	AUC	Sensibilité	Spécificité	VPP	VPN
Random Forests	76,6%	90,6%	41,6%	73,4%	71,3%
SVM	75,1%	95,2%	29,9%	70,5%	77,3%
XGBoost	74,0%	82,4%	52,8%	75,7%	62,8%
Régression logistique pénalisée	64,9%	58,8%	58,4%	71,6%	44,3%
Régression logistique	62,4%	57,8%	56,7%	70,4%	43,0%
Réseaux de neurones	62,0%	68,7%	44,2%	68,7%	44,2%
MARS	52,0%	99,3%	4,7%	65,0%	78,6%
Méthode d'ensemble	77,1%	92,5%	34,8%	71,6%	72,3%

3.3.4. Importance des variables prédictives

Les variables avec la plus grande importance dans les prédictions à cinq ans sont globalement les mêmes que dans les prédictions à deux ans. Parmi les différences, on observe que la distance entre le médecin pilote et le lieu du domicile est la variable avec la plus grande importance, suivie du poids de naissance, de la durée d'hospitalisation néonatale et de l'âge de la mère à la naissance (tableau 20). Parmi les 20 variables les plus prédictives on retrouve également toutes les dimensions cliniques néonatales (durée d'hospitalisation, poids de

naissance, terme de naissance...), cliniques maternelle ou liés à l'accouchement (âge de la mère, prématurité provoquée), socio-démographiques (niveau de vie et taux de pauvreté dans la commune de résidence), géographiques (distance), socio-économiques (situation professionnelle, niveau d'étude) et les dimensions propres au programme HYGIE-SEV (type de médecin) ou en lien avec la temporalité de prise en charge (semaine de naissance, année de naissance).

Tableau 20 : Extraction des 20 variables avec la plus grande mesure d'importance dans le modèle des *Random Forests* pour la prédiction de l'absence de visite à cinq ans, cohorte HYGIE-SEV, échantillon test (N = 648)

Variable	Valeur de la mesure d'importance
Distance	58
Poids de naissance	57
Durée d'hospitalisation néonatale	57
Age de la mère à la naissance	56
Terme de naissance	53
Poids de sortie	52
Semaine de naissance	51
Z score du poids de naissance	36
Niveau d'étude (supérieur)	30
Situation professionnelle de la mère (En activité)	27
Mode de garde envisagé (père ou mère au foyer)	27
Niveau de vie dans la commune de résidence	22
Taux de pauvreté dans la commune de résidence	21
Allaitement (Oui)	21
Année de naissance (2016)	20
Profession du conjoint (Employé)	19
Type de médecin (Libéral)	19
Sexe (Masculin)	19
Type de médecin (Hospitalier)	19
Profession du conjoint (Cadre ou profession intellectuelle supérieure)	19

3.4. Résultats du modèle incrémental

En suivant rigoureusement la même procédure que pour les analyses antérieures, il a été observé une amélioration notable des performances des modèles de prédiction à deux et à

cinq ans. Plus précisément, pour la méthode des *Random Forests* sur les données de test, l'AUC pour le modèle de deux ans a augmenté de 77,8 % à 85,4 % (85,2 % pour la méthode d'ensemble). De même, pour le modèle à cinq ans, l'AUC est passée de 77,8 % à 89,5 % (89,4 % pour la méthode d'ensemble). Quant aux autres modèles individuels, tous ont montré une amélioration de leurs performances à deux ans, tout en conservant l'ordre en termes d'AUC. A cinq ans, chaque modèle individuel a également vu ses performances améliorées.

En ce qui concerne l'importance des variables, les mêmes variables ont été retrouvées et les variables concernant la réalisation des précédentes visites étaient celles avec la plus haute valeur d'importance (résultats non montrés).

4. Discussion

4.1. Synthèse des résultats

En utilisant la méthode d'ensemble, une AUC supérieure à 75 % a été atteinte pour les prédictions à deux ans et à cinq ans, ce qui en fait un résultat prometteur. En comparant les différents modèles, celui des *Random Forests* s'est avéré le plus performant. Ce modèle présente l'avantage de fournir de bonnes prédictions tout en étant moins consommateur en termes de temps de calcul que d'autres modèles, comme le SVM. Les algorithmes de *Random Forests* et de SVM sont d'ailleurs ceux qui affichent les meilleures performances, surpassant la régression logistique. En revanche les performances des réseaux de neurones étaient inférieures aux modèles de régression logistique et de régression logistiques pénalisée. La méthode d'ensemble ne permettait pas d'augmenter les performances des résultats obtenus.

L'utilisation des *Random Forests* permet également de déterminer l'importance des variables pour identifier les facteurs prédictifs les plus pertinents à l'inclusion. À deux ans comme à cinq ans, les variables avec les plus hautes valeurs d'importance étaient les variables cliniques néonatales (terme de naissance, poids de naissance, durée d'hospitalisation) ou maternelles (âge de la mère à la naissance), géographiques ou démographiques (distance avec le lieu de suivi, niveau de vie et taux de pauvreté dans la commune de résidence), structurelles (semaine de naissance, type de médecin pilote) et encore socio-économiques (niveau d'étude, situation professionnelle des parents). Cela suggère que la prédiction repose sur un aspect multi-

dimensionnel et que les variables prédictives doivent être prises dans leur ensemble pour obtenir des résultats de prédiction satisfaisants. De plus, ces variables ont également été retrouvées comme facteurs associés à la perte de vue lors de l'étude associée du chapitre III, ce qui souligne la cohérence entre les analyses de données de HYGIE-SEV.

En incluant les variables de suivi, dans les modèles incrémentaux, une amélioration des résultats était notable. En effet, l'AUC atteignait alors 85,4 % pour les prédictions à deux ans et 89,5 % pour celles à cinq ans. Ces résultats soulignent l'importance de prendre en compte les données de suivi dans l'évaluation du modèle, car elles contribuent de manière significative à renforcer la précision des prédictions. Cette observation met en évidence le potentiel des *Random Forests* lorsqu'elles sont appliquées avec l'intégration des variables de suivi pertinentes, offrant une base solide pour de futures recherches et applications dans ce domaine. Concrètement cela se traduit par la prédiction de la réalisation d'une visite sur l'autre. Par exemple, prédire à quatre ans le fait de savoir si l'enfant fera ou non sa visite des cinq ans.

4.2. Comparaison avec la littérature

Les résultats sont globalement en accord avec ceux rapportés dans l'étude de Teixeira et coll. (205). En particulier, nous observons des performances améliorées avec l'utilisation de l'algorithme de *Random Forests*, une similitude dans l'importance attribuée aux variables, et une amélioration des performances avec l'application de la méthode d'ensemble. Cependant, par rapport à l'étude de Teixeira et coll. (205), certains axes d'amélioration sont notables. Tout d'abord, la taille de la population pour notre étude était considérablement plus grande que celle de l'étude de Teixeira et coll., avec plus de 9 000 enfants contre 542. Cela pourrait avoir une influence sur la robustesse de nos modèles et sur la confiance que nous avons dans nos résultats. En effet, une plus grande taille d'échantillon peut offrir une plus grande puissance statistique et permettre de détecter des effets plus subtils. De plus le nombre de variables disponibles dans HYGIE-SEV comparé à l'étude de Teixeira et coll. notamment en ce qui concerne les variables socio-économiques est nettement supérieure (24 variables contre 82 ici).

De plus, alors que les *Random Forests* ont montré de meilleures performances dans les deux études, il est important de noter que cet algorithme a également démontré une efficacité supérieure dans de nombreuses autres applications médicales (203,214,215). Cela renforce l'idée que cette méthode pourrait être particulièrement adaptée à la modélisation de données de santé complexes et multidimensionnelles.

En conclusion, non seulement nos résultats sont en accord avec ceux de l'étude de Teixeira et coll. (205) mais ils apportent une contribution supplémentaire en confirmant et en étendant ces résultats à un échantillon de patients beaucoup plus large.

4.3. Forces et limites de l'étude

Les premiers résultats obtenus dans cette étude sont prometteurs, avec une AUC supérieure à 75 %, ce qui indique une performance satisfaisante du modèle autant à deux ans qu'à cinq ans. L'un des principaux atouts de HYGIE-SEV réside dans l'enrichissement continu de ses données. En effet, les individus sont ajoutés année après année, ce qui accroît la puissance de calcul et, par conséquent, la qualité des résultats obtenus. De plus, la richesse des variables prédictives est un autre point fort de cette étude. Ces variables sont nombreuses et diversifiées, couvrant plusieurs dimensions telles que les aspects socio-économiques, cliniques et spécifiques aux programmes. Cette variété permet d'obtenir une vision plus complète et précise des facteurs influençant les résultats, ce qui renforce la robustesse et la pertinence des modèles développés. La combinaison de différentes méthodes de prédiction permet de tirer parti des avantages spécifiques de chaque technique et d'améliorer encore les performances de prédiction. Cette approche contribue également à la généralisation et à la flexibilité des modèles, ce qui est essentiel pour leur application dans divers scénarios et contextes. Dans l'ensemble, ces forces soutiennent le potentiel de HYGIE-SEV en tant qu'outil de recherche et de prédiction efficace et fiable, offrant des perspectives intéressantes pour les futures études et applications dans ce domaine. Une autre force de l'étude est l'effort fourni quant à l'optimisation des modèles. En effet, en combinant les étapes de pré-traitement, de validation croisée et d'optimisation des hyperparamètres, la méthodologie employée permet une certaine robustesse des résultats.

Il convient de souligner certaines limites qui pourraient influencer l'interprétation des résultats. Tout d'abord, malgré l'effort de généralisation de la méthode, il est possible que les résultats ne soient pas directement applicables à d'autres populations ou contextes, et des études supplémentaires pourraient être nécessaires pour valider les modèles dans des situations différentes. La qualité des données utilisées dans l'étude peut influencer les performances des modèles. Les erreurs de mesure, les données manquantes ou les biais inhérents aux données recueillies pourraient affecter les prédictions et les conclusions tirées. Il est essentiel d'évaluer et de minimiser ces facteurs pour garantir des résultats robustes et fiables. Bien que l'étude ait examiné plusieurs méthodes de prédiction et exploité la richesse des variables prédictives, il existe d'autres approches et techniques qui pourraient être explorées pour améliorer davantage les performances de prédiction. D'autres algorithmes de prédiction ou l'ajout de variables supplémentaires pourraient aider à résoudre certaines lacunes des modèles actuels et à affiner les prédictions. Enfin, la complexité et l'interprétation des modèles peuvent représenter un défi, en particulier pour les praticiens et les décideurs qui cherchent à utiliser les résultats dans la pratique.

4.4. Perspectives

En plus de l'enrichissement de nouvelles données par l'inclusion et le suivi de nouveaux enfants, dans les perspectives futures de cette étude, plusieurs pourraient être explorées afin d'améliorer davantage les modèles de prédiction et d'élargir leur applicabilité.

Le processus de prédiction ayant volontairement été laissé adaptable à différentes données et différents modèles, une piste intéressante d'amélioration serait de perfectionner ce processus en le spécifiant pour coller précisément aux données de cette étude, notamment en adaptant les étapes de pré-traitement de données et en les optimisant uniquement pour le cas de figure des données incluses dans l'étude de la prédiction des perdus de vue. Par exemple ici on a effectué une transformation Box-Cox sur toutes les données numériques mais on pourrait très bien n'appliquer cela que sur seulement les variables qui ne suivent pas une distribution normale. De même, les étapes de pré-traitement de données sont les mêmes quel que soit l'algorithme utilisé. Cependant, les hypothèses de départ ne sont pas les mêmes entre un algorithme paramétrique, non paramétrique ou semi paramétrique, notamment en

ce qui concerne la relation linéaire entre les variables prédictives et la variable de sortie. Ainsi le pré-traitement devrait être adapté à chaque algorithme pour optimiser leur qualité de prédiction. Enfin, dans les perspectives concernant le pré-traitement de données, il resterait à diminuer notre nombre de variables d'entrée. En effet nous avons inclus ici l'ensemble des 50 variables prédictives mais, pour plusieurs raisons, on pourrait sélectionner seulement une partie de ces variables pour améliorer le processus de prédiction. En effet cela permettrait 1) d'améliorer le temps de calcul du modèle, 2) de faciliter l'interprétation des résultats, et 3) de limiter la colinéarité entre les variables dans les modèles paramétriques. En s'appuyant sur les premiers résultats et notamment sur la valeur de mesure d'importance, on pourrait se limiter par exemple aux 10 ou 15 variables les plus prédictives.

Les réseaux de neurones ayant montré des performances décevantes (inférieures à la régression logistique), il conviendra à l'avenir d'améliorer la programmation de cet algorithme en utilisant d'autres paramétrisations, par exemple en utilisant plusieurs couches cachées contre une seule dans cette étude.

L'intégration de nouvelles variables prédictives alimentant le processus d'enrichissement de données pourrait améliorer les résultats obtenus, telles que des informations démographiques ou socio-environnementales propres au lieu de vie et au lieu de suivi de l'enfant.

En plus de prédire dès l'inclusion la perte de vue, il pourrait également être utile de prévoir une prédiction d'une visite à l'autre. Comme évoqué dans l'analyse de sensibilité présentée en annexe 4, l'inclusion des variables de suivi augmente considérablement la qualité des prédictions avec une AUC de près de 90 % sur la prédiction de la visite réalisée à cinq ans lorsqu'on inclut les données de suivi jusqu'à quatre ans. Dans l'analyse des modèles incrémentaux, seul le fait d'avoir fait ou non les visites précédentes était inclus comme tel dans le modèle. Mais en ajoutant des variables telles que la croissance de l'enfant, son développement moteur et cognitif ou encore sa scolarité à cinq ans par exemple, de nouvelles informations supplémentaires pourront être prises en compte pour la prédiction.

Pour augmenter au choix la sensibilité ou la spécificité, le seuil de classification de la prédiction peut également être modifié. Dans cette étude il était de 50 % c'est-à-dire que si la probabilité que l'enfant n'effectue pas la visite est de plus de 50 %, alors la visite sera

prédite comme non effectuée. Pour, par exemple, augmenter la sensibilité, on pourrait diminuer ce seuil de classification ce qui augmenterait le taux de Vrais Positifs. La définition de ce seuil est également un choix clinique à trancher avec les professionnels du terrain et les investigateurs de HYGIE-SEV.

Les efforts pour améliorer la communication et l'interprétation des résultats pour les praticiens et les décideurs pourraient faciliter l'adoption et l'application des modèles dans la pratique. La création de visualisations intuitives de type interface web, de rapports simplifiés et d'outils interactifs pourrait aider à rendre les résultats plus accessibles et compréhensibles pour les professionnels de HYGIE-SEV.

Il serait pertinent d'évaluer l'impact réel des modèles de prédiction sur les décisions cliniques et les résultats des patients. Des études prospectives pourraient être menées pour comparer les décisions basées sur les modèles développés avec les approches actuelles, afin de déterminer si l'utilisation des modèles de prédiction améliore réellement la prise en charge des patients et les résultats à long terme.

Enfin, il serait intéressant d'étendre l'analyse à d'autres variables de sortie pour évaluer la généralisation des résultats et des modèles développés. Ainsi, le travail présenté dans la section suivante porte sur la prédiction du devenir neurodéveloppemental des enfants grands prématurés.

CHAPITRE 5 – GENERALISATION DES ALGORITHMES : OPTIMALITE DU NEURODEVELOPPEMENT

Ce chapitre présente un dernier travail de thèse visant à valider et étendre les algorithmes précédemment développés à d'autres contextes, ici la prédiction du neurodéveloppement à l'âge de deux ans en utilisant l'algorithme basé sur les *Random Forests*.

1. Introduction

Comme mentionné précédemment, plusieurs études tendent à montrer que les enfants nés grands prématurés sont plus à risque de développer des anomalies de développement à l'adolescence et à l'âge adulte (44–51). La détection précoce de ces anomalies de développement permettrait aux professionnels de santé qui suivent ces enfants d'adapter individuellement leur suivi et les soins prodigués durant la petite enfance pour limiter l'impact de telles anomalies sur le quotidien des enfants et des familles (58,65–68). D'où la nécessité de mettre en place un suivi individualisé (69).

Cependant, encore peu d'études se concentrent sur la prédiction du neurodéveloppement dès la naissance. Un travail existant sur la prédiction de la réhospitalisation chez des enfants prématurés (203) rapportent des performances plutôt satisfaisantes avec une AUC de 64,8 % pour un modèle de *Random Forests* mais sans doute susceptibles d'être améliorées, ce qui encourage à poursuivre et tester sur des échantillons plus grands le travail de prédiction d'*outcomes* cliniques dans les cohortes d'enfants grands prématurés.

Toujours en se basant sur les données issues du programme HYGIE-SEV, ce travail a permis d'appliquer l'algorithme de prédiction obtenu dans l'étude antérieure sur les perdus de vue au neurodéveloppement des enfants grands prématurés. L'objectif était double puisqu'il s'agissait, dans un premier temps, d'évaluer les performances de notre algorithme pour un autre *outcome* que celui des perdus de vue pour tester sa généralisation et, dans un second temps, d'utiliser une approche innovante sur la détection précoce des anomalies de développement survenant chez les enfants grands prématurés à deux ans.

2. Méthodologie

Bien que la méthodologie employée ait été globalement la même que pour la précédente étude, il a fallu adapter certaines parties pour être en accord avec les objectifs de cette nouvelle étude.

2.1. Population d'étude

Comme précédemment, la population étudiée était celle des enfants grands prématurés (<33 SA) inclus entre le 1^{er} novembre 2015 et le 31 décembre 2022. Les enfants n'ayant pas effectué leur visite des deux ans ont cette fois été exclus des analyses quel qu'en soit le motif puisqu'il était alors impossible d'évaluer leur neurodéveloppement. Cependant, pour éviter un biais de sélection, les enfants sortis du programme HYGIE-SEV pour être pris en charge dans un programme de suivi spécialisé, ont été inclus dans les analyses et considérés comme ayant un neurodéveloppement non optimal à deux ans.

2.2. Définition de la variable de sortie

L'*outcome* étudié ici était l'optimalité du neurodéveloppement. Pour évaluer l'optimalité du neurodéveloppement chez les enfants à deux ans, deux questionnaires d'évaluation ont été utilisés (cf. tableau S4 en annexe 3). Le questionnaire standardisé ASQ a également été inclus parmi les critères d'évaluation. Par ailleurs, d'autres critères d'évaluation présents dans le

formulaire des deux ans ont aussi été utilisés. Ainsi, un enfant présentant au moins une des neuf conditions suivantes était considéré comme ayant un neurodéveloppement non optimal :

1. L'enfant ne marche pas, y compris avec de l'aide
2. Score moteur égal à 3 ou 4
3. Score de motricité fine égal à 3 ou 4
4. Détection d'une anomalie cognitive sévère ou légère
5. Difficultés motrices
6. Déficit du moteur spastique
7. Anomalie de l'axe
8. Trouble neurovisuel suspect ou avéré
9. Somme des items ASQ inférieure à un seuil fixé à 180

2.3. Modèle prédictif et optimisation

Dans les résultats précédents nous avons vu que l'algorithme de *Random Forests* était celui qui donnait les meilleurs résultats en terme d'AUC. Dans cette logique, c'est cet algorithme qui a été utilisé pour prédire le neurodéveloppement des enfants à deux ans. L'utilisation d'un seul algorithme plutôt que sept comme précédemment a l'avantage de permettre d'adapter le pré-traitement des données (la recette de *Tidymodels*) à cet algorithme. Notamment en se passant des étapes de transformation des variables continues telles que la transformation Box-Cox qui est plus pertinente pour les modèles paramétriques. Le choix s'est également porté sur les *Random Forests* pour limiter l'effet « boîte noire » lorsqu'on implémente plusieurs algorithmes afin de faciliter l'interprétation des modèles en se basant sur l'importance des variables.

La même méthode était utilisée pour la sélection des hyperparamètres avec la grille de recherche hypercube latin, pour le nombre de sous-ensembles en validation croisée et pour le choix de la métrique d'évaluation. Enfin la proportion du jeu d'entraînement était également de 80 % et les variables prédictives incluses dans le modèle étaient les mêmes que dans l'étude des perdus de vue.

2.4. Analyses statistiques

Les caractéristiques des enfants classés comme ayant un neurodéveloppement non-optimal ou optimal ont été décrites à l'aide d'effectifs et proportions pour les variables catégorielles, et de moyennes et écarts-types pour les variables continues.

3. Résultats

Pour cette étude, 6 546 enfants ont été inclus dans les analyses principales (figure 18). Parmi eux, 2 025 (30,9 %) enfants étaient identifiés comme ayant un neurodéveloppement non optimal. Les caractéristiques ayant servies à déterminer l'optimalité du neurodéveloppement sont détaillées en tableau 21. Il est à noter que, prise isolément, chaque caractéristique permettait de détecter une non-optimalité dans, au maximum 8,7 % des cas (pour l'anomalie cognitive). Cependant, en combinant l'ensemble de ces caractéristiques, 30,9 % des enfants présentaient une non-optimalité sur au moins un de ces critères. On constate aussi que pour les questionnaires standardisés propres à HYGIE-SEV (score moteur et de motricité fine), seulement 1,4 % et 2,0 % de données manquantes étaient observées contre 80,2 % pour le score standardisé validé (ASQ).

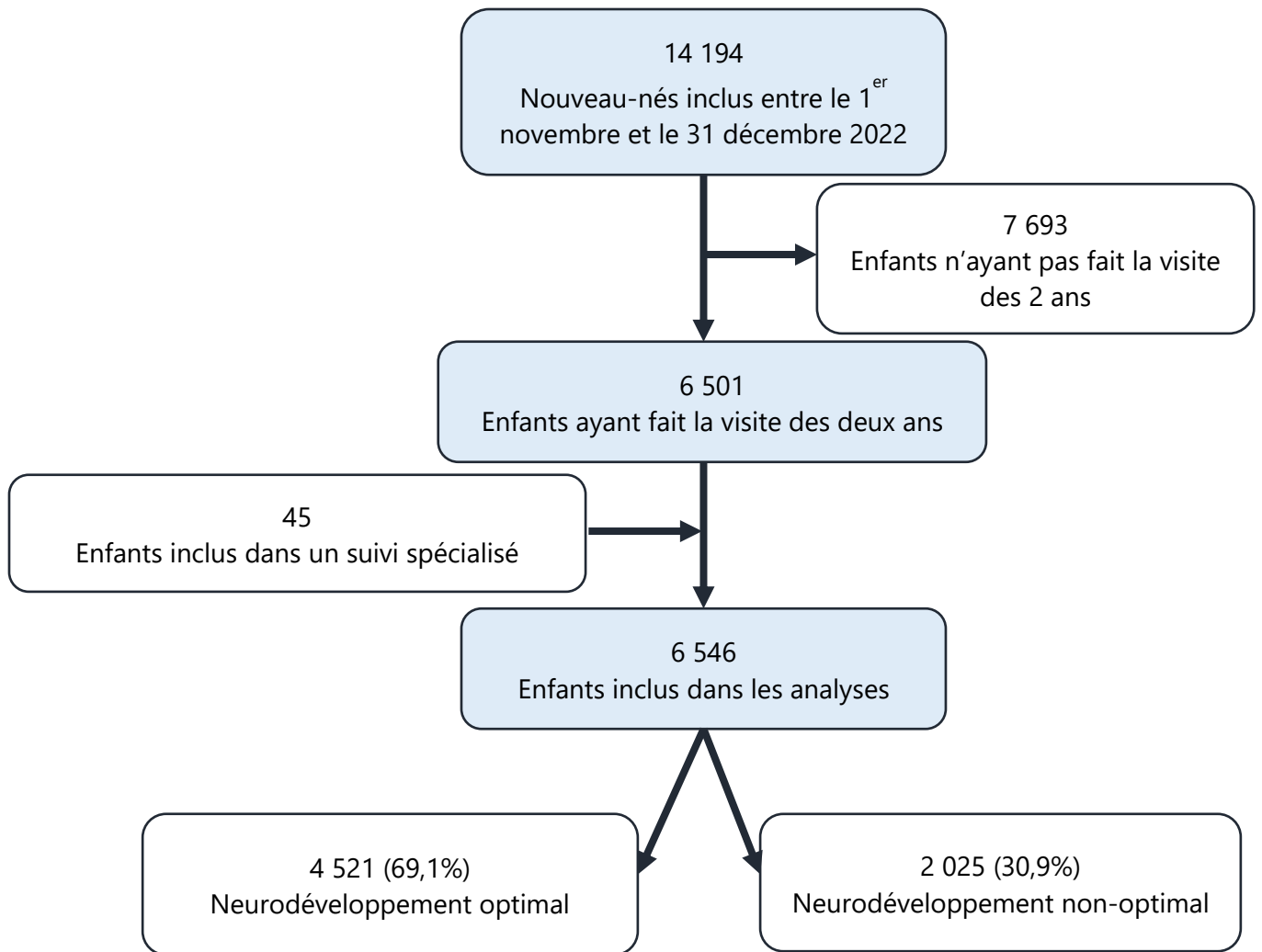


Figure 18 : Sélection de la population éligible aux analyses prédictives du neurodéveloppement, cohorte HYGIE-SEV

Tableau 21 : Principales caractéristiques des enfants sur les critères d'optimalité du neurodéveloppement à deux ans (N = 6 546)

Caractéristiques	n	%
Neurodéveloppement		
Non-optimal	2 025	30,9
Optimal	4 521	69,1
Marche		
Oui	5 634	86,1
Non	71	1,1
Inconnu	841	12,8
Score moteur		
1	6 011	91,8
2	293	4,5
3	109	1,7
4	39	0,6
Inconnu	94	1,4
Score de motricité fine		
1	5 511	84,2
2	748	11,4
3	123	1,9
4	33	0,5
Inconnu	131	2,0
Anomalie cognitive		
Oui	571	8,7
Non	4 600	70,3
Inconnu	1 375	21,0
Difficultés motrices		
Oui	306	4,7
Non	5 525	84,4
Inconnu	715	10,9
Déficit du moteur spastique		
Oui	123	1,9
Non	5264	80,4
Inconnu	1 159	17,7
Anomalie de l'axe		
Oui	126	1,9
Non	4 523	69,1
Inconnu	1 897	29,0
Trouble neurovisuel		
Oui	64	1,0
Non	3 398	51,9
Inconnu	3 084	47,1
ASQ		
Inférieur 180	148	2,3
Supérieur ou égal à 180	1145	17,5
Inconnu	5 253	80,2

Les principales caractéristiques des enfants ayant un neurodéveloppement optimal et non-optimal sont présentées dans le tableau 22. On semblait notamment retrouver une proportion plus importante de garçon chez les enfants ayant une non-optimalité de neurodéveloppement, des âges gestationnels plus faible également chez les enfants avec une non-optimalité du neurodéveloppement, de même qu'une durée d'hospitalisation plus importante et une proportion d'enfant avec une évaluation neurologique à risque élevée. On retrouve aussi des différences de répartition entre un médecin libéral et un médecin hospitalier mais aussi dans les départements de résidence et les réseaux d'inclusion, en particulier on retrouve 53,8 % d'enfants inclus dans le réseau RPSOF parmi les enfants ayant non-optimalité du neurodéveloppement contre 20,9 % parmi les enfants avec un neurodéveloppement optimal.

Tableau 22 : Principales caractéristiques des enfants selon l'optimalité du neurodéveloppement à deux ans (N = 6 546)

Caractéristique	Neurodéveloppement non-optimal (N = 2 025)		Neurodéveloppement optimal (N = 4 521)	
	n	%	n	%
Sexe de l'enfant				
Féminin	939	46,4	2 209	48,8
Masculin	1 086	53,6	2 312	51,2
Age gestationnel (semaines d'aménorrhée)				
23	8	0,4	9	0,2
24	76	3,8	100	2,2
25	142	7,0	221	4,9
26	163	8,0	277	6,1
27	214	10,6	357	7,9
28	175	8,6	461	10,2
29	225	11,1	585	12,9
30	287	14,2	684	15,1
31	338	16,7	756	16,7
32	397	19,6	1071	23,7
Evaluation neurologique				
Non fait	37	1,8	86	1,9
Normal ou bas risque	1 670	82,5	4 038	89,3
Risques modérés	228	11,3	349	7,7
Risques élevés	90	4,4	48	1,1
Réseau d'inclusion				
RPSOF	1 089	53,8	946	20,9
RSPP	295	14,6	925	20,5
NEF	222	11,0	1 181	26,1
RPVM	207	10,2	538	11,9
RPVO	79	3,9	486	10,7
MYPA	133	6,6	445	9,8
	Moyenne	Ecart-type	Moyenne	Ecart-type
Poids à la naissance (g)	1 237	412	1 293	391
Poids à la sortie (g)	2 661	551	2 603	512
Durée d'hospitalisation (jours)	70	43	62	36

Tableau 22 (suite)

Caractéristique	Neurodéveloppement non optimal (N = 2 025)		Neurodéveloppement optimal (N = 4 521)	
	n	%	n	%
Département de résidence				
75	165	8,1	647	14,3
77	234	11,6	602	13,3
78	169	8,3	493	10,9
91	474	23,4	365	8,1
92	399	19,7	402	8,9
93	210	10,4	874	19,3
94	256	12,6	598	13,2
95	96	4,7	477	10,6
Hors IDF	22	1,1	63	1,4
Type de médecin pilote				
Hospitalier	1 025	50,6	2 574	56,9
Libéral	640	31,6	1 118	24,7
Autre	146	7,2	382	8,4
Inconnu	214	10,6	447	9,9
Evaluation neurologique				
Non réalisée	37	1,8	86	1,9
Normale ou bas risque	1 670	82,5	4 038	89,3
Risques modérés	228	11,3	349	7,7
Risques élevés	90	4,4	48	1,1

Sur les données d'entraînement, l'optimisation des hyperparamètres a donné les résultats suivants présentés dans le tableau 23.

Tableau 23 : Résultats de l'optimisation des hyperparamètres dans le modèle de prédiction du neurodéveloppement à deux ans, cohorte HYGIE-SEV, échantillon d'entraînement (N = 5 236)

Modèle	Hyperparamètre	Valeur
Random Forests	ntrees	1 526
Random Forests	mtry	15
Random Forests	min_n	2

L'AUC moyenne établie par validation croisée sur les données d'entraînement avec ces hyperparamètres était de 90,7 % (tableau 24).

Tableau 24 : Résultats des métriques des *Random Forests* sur les données d'entraînement.

Métrique	Moyenne	Ecart-type
AUC	90,7%	0,004
Sensibilité	76,8%	0,009
Spécificité	86,9%	0,001

L'ajustement du modèle a ensuite été établi sur l'entièreté du jeu de données d'entraînement. L'évaluation sur les données de test a montré une valeur d'AUC de 78,7 % (tableau 25).

Tableau 25 : Evaluation de la performance du modèle de *Random Forests* pour la prédiction du neurodéveloppement à deux ans, cohorte HYGIE-SEV, échantillon de test (N = 1 310)

Modèle	AUC	Sensibilité	Spécificité	VPP	VPN
Random Forests	78,7%	44,0%	89,0%	64,0%	78,0%

Enfin parmi les variables prédictives avec la plus grande valeur d'importance (tableau 26), on retrouve entre autre le réseau d'inclusion RPSOF, la durée de l'hospitalisation néonatale initiale ou encore la distances entre le lieu de domicile et le médecin pilote. On retrouve également les poids de naissance et de sortie, ainsi que l'âge gestationnel. L'âge de la mère à la naissance ou la semaine de naissance (variable temporelle) se retrouve aussi parmi les variables les plus prédictives. Le département de résidence (situé dans le réseau RPSOF) correspond à la 11^{ème} variable avec la valeur d'importance la plus haute. Des variables socio-économiques sont parmi les 20 variables les plus prédictives avec le mode de garde de l'enfant (père ou mère au foyer), le niveau d'étude de la mère (supérieur), le nombre d'enfants vivant au foyer et la profession du conjoint (employé).

Tableau 26 : Extraction des 20 variables avec la plus grande mesure d'importance dans le modèle des Random Forests pour la prédiction de d'un neurodéveloppement non-optimal, cohorte HYGIE-SEV, échantillon test (N = 1 310)

Variable	Valeur de la mesure d'importance
Réseau d'inclusion (RPSOF)	117
Durée d'hospitalisation	93
Distances	92
Poids de naissance	89
Terme de naissance	87
Poids de sortie	87
Semaine de naissance	83
Âge de la mère à la naissance	81
Z-score du poids de naissance	56
Sexe (Masculin)	56
Département de résidence (92)	49
Mode de garde (Père ou mère au foyer)	46
Niveau d'étude de la mère (Supérieur)	45
Allaitement (Oui)	44
Type de naissance (Unique)	44
Département de naissance (91)	43
Nombre d'enfant au foyer avant la naissance (2 ou plus)	41
Evaluation neurologique (Normal ou bas risque)	40
Nombre d'enfant au foyer avant la naissance (1)	40
Profession du conjoint (Employé)	36

4. Discussion

4.1. Synthèse des résultats

Cette première étude sur la prédiction du neurodéveloppement a mis en évidence un taux de non optimalité de 30,9 % à deux ans. D'après les résultats des algorithmes de prédiction, une AUC de 78,7 % a été trouvée sur les données de test grâce aux *Random Forests*. Ces premiers résultats sont prometteurs quant à la mise en place d'un algorithme validé permettant de prédire le neurodéveloppement des enfants à deux ans. La diversité des variables avec les valeurs d'importance les plus élevées soulignent l'importance de prendre en compte l'aspect

multidimensionnel du développement des enfants. De plus il n'est pas surprenant de constater que le réseau RPSOF soit parmi les variables les plus prédictives car la proportion d'enfants identifiés comme ayant un neurodéveloppement non-optimal à deux ans varie selon le réseau de suivi de l'enfant, le réseau RPSOF étant celui devançant de loin les autres réseaux (tableau 22).

4.2. Comparaison avec la littérature

La proportion de 30,9 % de neurodéveloppement non optimal était comparable à d'autres cohortes françaises similaires. Dans la cohorte LIFT (76), la proportion d'enfants nés avant 34 SA ayant un trouble du neurodéveloppement à deux ans était de 32,3 %. Dans une autre cohorte de recherche française, EPIPAGE-2 (187), ce taux était de 28,8 % pour les enfants nés avant 34 SA en ce qui concerne le développement neuromoteur ou la survenue d'anomalie sensorielle. Bien que les définitions soient variables d'une cohorte à l'autre, que la population étudiée ne soit pas exactement la même (<34 SA contre <33 SA ici), les proportions d'enfants ayant un neurodéveloppement non optimal à deux ans sont globalement similaires.

Dans la cohorte EPIPAGE-2, une étude s'est intéressée à la prédiction de la réhospitalisation des enfants (203) : sur les 3 841 enfants nés avant 34 SA inclus dans l'étude, 75 variables prédictives ont été prises en compte dans des modèles de prédiction incluant des *Random Forests*. Les résultats ont montré une AUC de 64,8 %. Cette valeur inférieure à la nôtre peut résulter de la réhospitalisation comme critère de jugement qui est généralement associé à de moins bonnes performances (200). En comparaison, les performances de notre modèle apparaissent donc de bonne qualité.

4.3. Comparaison avec l'étude des perdus de vue

En comparaison avec les résultats obtenus dans les modèles de prédiction de la perte de suivi, nous constatons des performances légèrement améliorées pour les prédictions du neurodéveloppement à deux ans. En effet, pour le modèle de *Random Forests*, l'AUC était de 78,4 % dans les prédictions des perdus de vue contre 78,4 % pour le neurodéveloppement.

Cette amélioration des résultats peut être due à plusieurs paramètres. En premier lieu, nous avons adapté la méthodologie et notamment le pré-traitement des données à un contexte plus propice aux *Random Forests*. Cependant cette amélioration étant limitée, elle mériterait d'être démontrée dans des études ultérieures pour confirmer que le neurodéveloppement à deux ans pourrait être plus prévisible que la perte de vue.

4.4. Forces et limites de l'étude

Les principales forces de l'étude présentée ici sont la taille de la population et son aspect « vie réelle ». C'est à ce jour, à notre connaissance, la plus grande cohorte de vie réelle qui s'intéresse à l'utilisation d'algorithmes de prédiction pour prédire le neurodéveloppement des enfants grands prématurés. Cette approche s'inscrit comme une innovation en termes de santé en périnatalité.

La méthodologie employée reste ici aussi un atout de l'étude puisqu'elle fait appel à plusieurs processus (grille de recherche, validation croisée, pré-traitement des données, imputation) permettant de rendre les résultats robustes.

L'utilisation de l'algorithme des *Random Forests* seul permet également une simplification de la méthodologie et une meilleure adaptation du pré-traitement de données à cet algorithme en plus de gagner du temps de calcul.

La principale limite de cette étude est la définition du neurodéveloppement. En effet, nous avons utilisé une définition arbitraire en utilisant les questionnaires d'évaluation disponibles dans HYGIE-SEV. Il serait également nécessaire de valider la construction de ce score de neurodéveloppement par des cliniciens. En l'état, cette définition n'a pas de validation externe et ne permet pas l'application directe de l'algorithme à d'autres contextes. Cependant, les résultats obtenus (AUC supérieure à 78 %) montrent que la définition utilisée ici pour l'optimalité du neurodéveloppement peut être une bonne piste à poursuivre dans HYGIE-SEV. En revanche, d'autres méthodologies devront être utilisées pour permettre la validation de cette définition à d'autres contextes.

La deuxième limite de l'étude est un potentiel biais de sélection de la population. En effet, sont exclus tous les enfants n'ayant pas effectué la visite des deux ans. Même si, comme vu

dans le chapitre 3, les enfants ayant des caractéristiques cliniques plutôt favorables étaient davantage susceptibles de ne pas réaliser les visites de suivi, il est possible qu'un certain nombre d'entre eux aient au contraire un neurodéveloppement non optimal. Ce sont donc autant d'informations manquées permettant d'améliorer les modèles. De même la sous-représentation des enfants pris en charge dans un suivi spécialisé est également un manque d'information disponible quant au neurodéveloppement non optimal. Cependant l'échantillon sur lequel cette étude s'est basée est suffisamment grand pour pouvoir avoir une bonne quantité d'information disponible.

Enfin nous avons également souligné (tableau 22) les différences de pratiques entre les réseaux de santé car on retrouve des proportions très différentes d'enfants ayant des anomalies de neurodéveloppement détectées à l'âge de deux ans. Les travaux entrepris par le conseil scientifique se penchent déjà sur cette problématique et optent pour une politique de régionalisation et de standardisation des pratiques.

4.5. Perspectives

Dans la perspective des travaux futurs, l'une des priorités sera de valider la pertinence d'optimalité du neurodéveloppement. En collaboration avec les professionnels de santé et le conseil scientifique, il conviendra d'affiner la définition du critère d'optimalité du neurodéveloppement. Pour l'approbation de cet algorithme, il est également nécessaire de l'implémenter sur de nouvelles données, afin d'évaluer a posteriori sa qualité et sa précision sur les enfants qui effectueront plus tard leur visite des deux ans.

Comme c'est le cas pour les prédictions des enfants perdus de vue, il sera possible de moduler le seuil de classification pour équilibrer de manière optimale sensibilité et spécificité, en fonction des priorités de l'étude. La prédiction du neurodéveloppement à deux ans représente une estimation à moyen terme, mais il serait également pertinent d'évaluer l'efficacité de notre algorithme pour des prédictions à plus long terme, notamment en établissant un score de neurodéveloppement à cinq ans, et d'appliquer notre modèle à deux ans pour évaluer les capacités prédictives.

Enfin, dans les perspectives plus générales, nous suggérons de mettre en place une homogénéisation des pratiques notamment entre les réseaux car nos résultats laissent entrevoir des différences notables entre les professionnels ce qui a comme répercussions directes des inégalités concernant le diagnostic du neurodéveloppement non optimal.

CHAPITRE 6 –

DISCUSSION GENERALE

1. Synthèse des résultats obtenus

1.1. Le système d'information HYGIE-SEV

Face à la complexité du système d'information lié au dispositif HYGIE-SEV et à la richesse des données disponibles, il a fallu, dans ce travail de thèse, prendre en compte un nombre de paramètres important pour faciliter les analyses tout en s'appuyant sur les ressources informatiques à disposition pour réaliser ces analyses.

Afin de prendre en compte les nouvelles inclusions, les ajouts ou modifications de formulaires (visites, sortie de réseau, adresses, etc.) et d'être aussi à jour que possible, il est indispensable d'adopter une démarche d'adaptation et de mise à jour constante. Concrètement cela passe par une automatisation de tous les programmes de création de tables et de *data management*. Le package *CronR* (216) permet de programmer l'exécution récurrente d'un script R. Dans le cadre de HYGIE-SEV, une récurrence journalière a été programmée pour alimenter les tables d'analyse de la cohorte. Cependant, pour les questions de recherche abordées dans les travaux de thèse présentés ici, il a fallu geler les bases à un instant t pour garantir la reproductibilité des résultats lors de nouvelles exécutions des scripts construits lors des différentes analyses.

L'importance d'utiliser des ensembles de données riches, à la fois en termes de nombre de variables et de nombre d'individus, est un point crucial dans le développement des modèles prédictifs. L'utilisation de nombreuses variables offre une compréhension plus détaillée et nuancée des facteurs qui influencent les performances des résultats des modèles de prédiction (217). De plus, des échantillons de taille importante renforcent la robustesse et la fiabilité des résultats obtenus, en garantissant une meilleure représentativité de la population étudiée et en augmentant la puissance statistique pour identifier des facteurs associés.

L'utilisation de données réelles et longitudinales représente une grande force de cette étude. Cela permet de suivre les individus au fil du temps et d'observer les évolutions et les tendances, ce qui est essentiel pour comprendre les phénomènes épidémiologiques. Ainsi, grâce au constant enrichissement des données et à l'automatisation de l'alimentation de la base de données, les résultats descriptifs et statistiques ne sont pas figés dans le temps mais représentent bien une réalité de terrain dynamique, ce qui en constitue un atout majeur pour les professionnels de santé participant au programme HYGIE-SEV qui peuvent ainsi avoir des résultats de la cohorte en temps réel. La grande diversité des variables disponibles dans le système d'information est également un atout de cette étude car nous avons pu voir que l'aspect multidimensionnel était nécessaire lorsqu'on s'intéresse aux questions du parcours de soin et du neurodéveloppement dans les cohortes pédiatriques (66,71–75,79,80,178–181).

Dans la cohorte HYGIE-SEV, plusieurs questionnaires standardisés sont utilisés pour évaluer le neurodéveloppement des enfants. Cependant, un seul des questionnaires utilisés à deux ans, l'ASQ, a fait l'objet d'une validation externe rigoureuse (70). Cela peut poser problème quant à la validité externe des résultats obtenus car les résultats ne sont potentiellement pas répliquables dans d'autres cohortes d'enfants vulnérables. La force majeure des scores disponibles dans HYGIE-SEV réside néanmoins dans leur très grande simplicité d'utilisation comparativement aux questionnaires validés qui font référence dans le domaine. De plus, dans HYGIE-SEV les médecins sont régulièrement formés à leur utilisation. Ainsi, contrairement à l'ASQ, les taux de remplissage sont nettement plus satisfaisants : environ 1 ou 2 % de données manquantes pour la majorité des autres scores à deux ans contre près de 80 % pour l'ASQ (tableau 21).

Même si les enjeux semblent a priori différents entre les cohortes de recherche épidémiologique et les programmes de suivi clinique (tel que HYGI-SEV), il est important de les mettre en perspective pour comprendre les mécanismes qui entrent en jeu notamment dans la perte de vue dans HYGIE-SEV. Le nombre élevé d'enfants perdus de vue est ainsi la principale limite des études longitudinales réalisées à partir de HYGIE-SEV. En effet, en comparant à d'autres cohortes similaires, nous avons pu voir que le taux de perdus de vue était plus élevé dans le programme HYGIE-SEV (73,74,76,77). Pour limiter au maximum ce biais, l'analyse des facteurs associés à la perte de vue est une nécessité dans cette cohorte. Dans une perspective de santé publique, cette analyse est aussi indispensable pour identifier

le plus précocement possible les enfants susceptibles d'être perdus de vue. Ceci dans le but de développer des stratégies de rétention efficaces permettant d'assurer un suivi optimal à la plus grande majorité des enfants vulnérables.

Malgré les limites du programme HYGIE-SEV, la richesse des données contenues dans son système d'information, son mode d'inclusion au fil de l'eau, son large bassin de recrutement (région Ile-de-France) et son infrastructure de traitement permettant une analyse en temps réel des données en font une cohorte unique pour répondre à de multiples problématiques de santé publique concernant les enfants vulnérables.

1.2. Complexité de la question des perdus de vue

L'étude des perdus de vue dans la cohorte HYGIE-SEV fut la première étude épidémiologique réalisée sur les données issues du programme HYGIE-SEV. Cette question est primordiale d'un point de vue de recherche et de santé publique pour, d'une part, comprendre pourquoi un nombre important d'enfants inclus dans le programme HYGIE-SEV sont perdus de vue et, d'autre part, enrichir les connaissances scientifiques sur le sujet des perdus de vue dans les cohortes d'enfants vulnérables en les confrontant aux résultats de la littérature. Les travaux de cette thèse présentés au chapitre 3 ont pu permettre d'ébaucher des premiers éléments de réponse à ces différentes problématiques.

Les résultats obtenus dans l'étude des facteurs associés à la perte de vue illustrent bien le réel enjeu autour de cette question dans le programme HYGIE-SEV. La proportion d'enfants perdus de vue est en effet plus importante que celles rapportées dans d'autres cohortes dans la littérature. Contrairement à HYGIE-SEV, ces cohortes sont souvent des cohortes de recherche dont le recrutement est limité dans le temps et pour lesquelles une attention particulière est portée sur la qualité des données tout au long du suivi. Cette attention vise à minimiser les biais liés à la qualité des données lors des analyses subséquentes. HYGIE-SEV est quant à elle une cohorte opérationnelle avec un recrutement au fil de l'eau et dont les données sont issues du dossier patient électronique. Dans HYGIE-SEV, le but est avant tout de limiter le nombre de perdus de vue pour assurer un suivi optimal au plus grand nombre d'enfants vulnérables. L'étude présentée dans cette thèse fournit des premiers éléments

basés sur les preuves nécessaires à la mise en place de programmes spécifiques « *evidence-based* » visant à améliorer la rétention des patients dans le programme HYGIE-SEV. Cette question est un axe de recherche prioritaire du conseil scientifique depuis sa création en 2020. Les travaux réalisés dans le cadre de cette thèse lui ont ainsi permis d'initier des premières actions. Par exemple, à partir de juin 2023, le conseil scientifique de HYGIE-SEV a pris la décision de mettre en place une mesure de relance par mail. Des emails automatiques sont ainsi envoyés aux familles deux mois avant la visite préprogrammée des deux ans. Dans ce mail, il est rappelé aux familles qui n'ont pas encore pris rendez-vous de se rapprocher du secrétariat médical pour en fixer la date. Pour les familles qui ont déjà fixé le rendez-vous, ce mail sert de simple rappel à la visite. L'impact de ce programme sur la période de juin à décembre 2023 sera évalué a posteriori.

Nos analyses ont permis d'identifier des caractéristiques à la fois cliniques, socio-économiques et concernant la prise en charge comme étant associées significativement à la non réalisation des visites à deux et cinq ans. De nombreux facteurs cliniques néonataux ont déjà été décrits dans d'autres cohortes comme étant associés à la perte de vue à deux ans comme à cinq ans (66,72,77,79,80,180). Dans le programme HYGIE-SEV, comme dans ces autres cohortes, les facteurs cliniquement favorables sont associés positivement à la perte de vue. Une explication possible pourrait être que les médecins qui suivent les enfants, tout comme les familles, prêtent davantage attention au suivi des enfants avec des problèmes de développement plus importants. D'un point de vue médical, cela reste problématique car les recommandations nationales (159,160,218) et internationales (69) préconisent explicitement le suivi jusqu'à l'âge de sept ans. Bien que ces enfants se développent a priori de manière plus favorable, ils restent néanmoins plus exposés à des anomalies de développement durant le reste de leur enfance, à l'adolescence et l'âge adulte que des enfants nés à terme (65–67).

Les facteurs cliniques ne sont pas les seuls associés à la perte de vue dans le programme HYGIE-SEV. Les résultats ont également montré que, à état néonatal égal, les enfants issus de familles ayant des caractéristiques socio-économiques défavorables sont davantage susceptibles de manquer leur visite. Le détail de ces caractéristiques est explicité dans le chapitre 3, mais celles-ci sont en accord avec la littérature (71,73,75,77,79). Les difficultés sociales de ces familles semblent se répercuter sur la qualité du suivi dans le programme HYGIE-SEV. Les hypothèses pour expliquer cette association sont nombreuses

(71,73,75,77,79). Cela a directement pour conséquence la perte de suivi d'enfants avec une double fragilité à la fois clinique et sociale. Cette double peine suggère la nécessité d'une attention particulière des professionnels de santé concernant cette population à risque. Dans une optique de médecine personnalisée, ces résultats sont un premier support à la définition de nouvelles stratégies de rétention focalisées sur cette population (adaptation du parcours de soins et de l'accompagnement au contexte social).

La troisième dimension associée à la perte de vue concerne les différences de suivi selon les caractéristiques générales du programme HYGIE-SEV. L'hétérogénéité du suivi selon le type de médecin pilote, les territoires de suivi ou encore le niveau de soins de la maternité de naissance reflète l'importance du contexte dans lequel l'enfant est inclus et suivi dans le programme. Des pistes de travail ont déjà été amorcées par le conseil scientifique, notamment en ce qui concerne l'uniformisation des pratiques sur les différents réseaux de périnatalité. Certains territoires d'Ile-de-France subissent depuis plusieurs années une démographie médicale et paramédicale insuffisante (219). Les résultats observés dans ces travaux de thèse illustrent bien ces difficultés rencontrées sur le terrain.

1.3. Des pistes d'amélioration pour le suivi personnalisé

Les travaux initiés ici sur le traitement des données massives constituent une approche innovante en santé pédiatrique. Il n'existe encore que peu d'études qui abordent la question de la médecine personnalisée en utilisant des algorithmes de prédiction dans les cohortes d'enfants vulnérables. Les algorithmes de prédiction élaborés au cours de cette thèse pourraient représenter des pistes d'exploration pour de futurs travaux de recherche qui seront menés au sein du programme HYGIE-SEV. Ces initiatives constitueraient un pas en avant vers une compréhension plus approfondie et un traitement plus efficace des problématiques de santé chez les enfants. D'une part, les travaux qui ont porté sur les perdus de vue serviront à la vigilance des médecins pilotes avec la construction d'un score de risque d'être perdu de vue. La perspective de prédire le devenir des enfants vulnérables à court (deux ans) et moyen terme (cinq ans) est ainsi un atout de la cohorte HYGIE-SEV. D'autre part, en s'appuyant sur la méthodologie développée au cours de ces premiers travaux, les prédictions pourront être déclinées à d'autres variables de sortie, comme le neurodéveloppement. Dans le chapitre 5,

les premiers résultats ont été présentés en utilisant la méthode des *Random Forests*. Avec une performance à plus de 78 % d'AUC, ces premiers résultats sont prometteurs. Dans le cadre de la médecine personnalisée, la mesure de l'importance des variables autorise également les investigateurs et les médecins pilotes de HYGIE-SEV à se focaliser plus précisément sur des aspects particuliers sur lesquels s'appuyer pour anticiper la survenue de la perte de vue ou d'anomalies du neurodéveloppement. On retrouve entre autres des variables d'ordre clinique comme la durée d'hospitalisation, le terme de naissance ou le poids de naissance mais aussi d'autres variables telles que l'âge de la mère ou la distance entre le lieu du domicile et le lieu de suivi.

Bien que ces travaux en soient encore à leurs débuts, et malgré certaines limites évoquées dans les chapitres 4 et 5, ils constituent une première ébauche de l'influence de la médecine personnalisée sur la santé pédiatrique de par leurs aspects innovants mais aussi leurs aspects pratiques car ils peuvent devenir de vrais outils d'aide à la décision pour les professionnels de santé. Avec le développement de l'IA à travers l'utilisation d'algorithmes de prédiction de toutes sortes, ce sont autant d'outils qui permettront de contribuer à améliorer le parcours de soins des enfants nés grands prématurés et de très faible poids de naissance, sans pour autant négliger le consentement du patient et l'impact du changement de la relation patient-médecin due à la généralisation de la médecine personnalisée (220).

2. Perspectives

2.1. Enrichissement des données pour compléter le parcours de soins

Pour mieux comprendre les raisons de l'attrition en fonction des variables contextuelles dans HYGIE-SEV (réseau de suivi, type de médecin pilote...), il serait intéressant d'enrichir les données d'autres variables représentant l'offre de soins du territoire francilien de manière générale mais aussi l'offre de soins disponible dans HYGIE-SEV. En plus des variables déjà disponibles, des informations concernant par exemple la démographie médicale par territoire seraient probablement une plus-value pour l'enrichissement de données. Après discussion avec les professionnels de santé, il apparaît que certains territoires de la région sont

dépourvus de ressources humaines suffisantes, notamment de médecins pilotes, pour suivre les enfants dans le programme HYGIE-SEV. Des données supplémentaires concrètes sur l'offre de soins pour étayer et confirmer ces propos seraient ainsi intéressantes pour enrichir le système d'information de HYGIE-SEV. De même, ce seraient des occasions supplémentaires d'augmenter le nombre de variables d'entrée dans les jeux de données utilisés pour réaliser les prédictions, ce qui pourrait permettre d'améliorer les performances des algorithmes. Dans la même logique, des données sur les caractéristiques socio-démographiques pourraient également être mises en perspective avec le reste des données utilisées pour compléter les résultats obtenus. Par exemple, au lieu des distances géographiques, il serait probablement plus intéressant de calculer les temps de trajet en transport (voiture, transport en commun, vélo) pour se rendre aux visites. Enfin, pour compléter le parcours de soins, il serait utile de collecter des données antérieures à la naissance de l'enfant pour ce qui concerne les caractéristiques cliniques et les antécédents de la mère. Pour cela, un appariement avec les données du Système National des Données de Santé (SNDS) (221) est également prévu à terme. L'appariement probabiliste entre HYGIE-SEV et le SNDS entre dans le cadre d'un projet plus large qui consiste à établir un parcours de soins complet de la mère, du nouveau-né et de leur devenir.

Comme il a été discuté précédemment, un programme de relance automatique a été mis en place en juin 2023. Ce programme a pour but d'augmenter l'exhaustivité des visites dans HYGIE-SEV afin de limiter l'attrition. Pour évaluer l'impact de ces relances, une étude sera mise en place a posteriori. Le but de cette étude sera de déterminer les profils d'enfants ou de familles pour lesquels les relances ont eu un impact significatif sur l'ensemble de la période couverte (de juin à décembre 2023).

Au rythme de 2 000 enfants grands prématurés ou de faible poids de naissance inclus par an, le nombre d'individus éligibles à la visite des cinq ans est estimé à environ 6 000 enfants en 2024 et pourra atteindre les 10 000 individus en 2026. Concernant l'éligibilité à la visite à deux ans, le nombre d'enfants devrait avoisiner les 20 000 d'ici cinq ans. Cette augmentation constitue une perspective naturelle dans le programme HYGIE-SEV et en fera une des plus importantes cohortes nationales de suivi d'enfants vulnérables. Malgré ce nombre important et la difficulté de traitement de données que cela peut causer, le travail réalisé dans cette

thèse sur la structuration des données et l'automatisation des programmes permettra de valoriser facilement les données du système d'information HYGIE-SEV.

2.2. Perspectives d'amélioration des algorithmes

Par souci de simplification, les étapes de pré-traitement ont été généralisées à l'ensemble des algorithmes de prédiction utilisés. Autrement dit, la procédure de pré-traitement des données était la même quel que soit l'algorithme utilisé. Afin d'optimiser les étapes de pré-traitement, il serait préférable d'adapter ces étapes en fonction de l'algorithme utilisé. Par exemple, certains algorithmes sont plus sensibles que d'autres à la transformation des données et adapter le pré-traitement à chaque algorithme pourrait améliorer la robustesse des résultats.

Bien que l'emploi de la grille de recherche hypercube latin soit d'un appui certain pour l'amélioration des performances des modèles utilisés, l'utilisation d'une technique d'optimisation bayésienne (222) pourrait contribuer à l'industrialisation des programmes de prédiction, notamment en ce qui concerne les temps de calcul. En effet, cette approche ne nécessite plus de définir une grille de recherche exhaustive et d'évaluer chaque combinaison d'hyperparamètres. À la place, l'optimisation bayésienne adapte et ajuste de manière itérative chaque hyperparamètre individuellement, en se basant sur les résultats obtenus lors des itérations précédentes. Cela permet de cibler plus efficacement l'espace des hyperparamètres, rendant le processus d'optimisation plus rapide et plus efficace.

L'exploration d'autres approches de prédiction pourrait également contribuer à l'amélioration des algorithmes proposés. Bien que les *Random Forests* semblent donner les meilleures performances, d'autres méthodes n'ont pas encore été explorées telles que la méthode de *Categorical* ou *Boosting catboost* (223) qui est une méthode utilisant le *gradient boosting* et qui permet, contrairement aux autres algorithmes utilisés ici, de gérer directement les données catégorielles sans passer par la méthode de *one-hot encoding*. De même la méthode des réseaux de neurones utilisée ici se limite à une seule couche intermédiaire entre la couche d'entrée et celle de sortie. Ajuster ce paramètre présente également une perspective d'amélioration des algorithmes. D'autres type de réseaux de neurones comme les réseaux *Convolutional Neural Networks* (CNNs) (224) pourraient être explorés pour améliorer les performances de ce type de modèle. Les CNNs se distinguent d'un réseau de neurones classique en limitant les connexions entre neurones d'une couche à l'autre.

2.3. Généralisation des algorithmes et création d'un outil à destination des professionnels

Le travail sur la prédiction du neurodéveloppement est un premier exemple sur lequel appliquer les modèles obtenus à d'autres contextes que les perdus de vue. Les programmes d'analyse ont été préparés de façon à faciliter leur adaptation, en automatisant les étapes de pré-traitement de données, en séparant les étapes d'application des algorithmes en fonctions, en parallélisant les calculs ou encore en définissant comme paramètres ajustables un certain nombre de choix du pré-traitement (nombre de sous-échantillons pour la validation croisée, modèles à utiliser, taille des grilles de recherche, nombre de cœurs à utiliser pour la parallélisation...). Ainsi, bien qu'initialement le travail ait été réalisé dans le but de prédire les perdus de vue, il a été relativement aisé d'appliquer les modèles à la problématique du neurodéveloppement.

Dans les perspectives à venir, il reste également la restitution des résultats aux professionnels de santé qui suivent les enfants inclus dans HYGIE-SEV, le but étant de mettre à disposition facilement les prédictions obtenues aux médecins. L'idée serait d'alerter dès l'inclusion si l'algorithme prévoit une perte de vue de l'enfant en fonction des caractéristiques de naissance et néonatales. Ainsi le médecin pourra être d'autant plus attentif au suivi du patient. De même, si l'algorithme prévoit une anomalie du neurodéveloppement, il serait important d'en notifier le médecin pilote.

Les médecins pilotes sont les premiers à pouvoir intervenir sur le maintien des enfants dans le programme HYGIE-SEV. Il est donc primordial de les convaincre de l'intérêt de ce programme. En effet, le remplissage des données à travers le logiciel HYGIE peut être vu comme une contrainte et une perte de temps surtout si aucun retour n'est effectué sur ces données remplies. En restituant et résumant toutes les informations saisies par les médecins, on leur montre ce qu'ils renseignent dans le logiciel et on les motive à continuer le suivi dans HYGIE-SEV.

Le développement d'une interface web à l'aide du package *Shiny* (225) a été initiée. En plus d'une synthèse globale des données issues de HYGIE-SEV (caractéristiques socio-économiques des familles incluses, caractéristiques néonatales des enfants, répartition des

médecin pilotes...), cette interface permettra de restituer aux médecins les résultats des prédictions obtenus grâce aux modèles présentés ici. En accédant via une connexion sécurisée, ils auront à disposition la liste de leurs patients et, en sélectionnant un des enfants, ils auront accès à un chronogramme ou *timeline* qui présentera les résultats des prédictions (par exemple pour chaque visite, une probabilité de ne pas réaliser la visite, en laissant le libre choix aux médecins de déterminer le seuil de classification). Rendre cette interface web doublement utilisable par, d'une part, l'ARS-IDF et, d'autre part, les professionnels de santé utilisant HYGIE-SEV et qui suivent des enfants permettrait de mettre en lumière l'ensemble des travaux produits à travers cette thèse CIFRe à cheval entre les questions de recherche et la problématique industrielle.

3. Conclusion

Cette thèse confirme l'importance du suivi des enfants vulnérables en particulier dans un contexte de vie réelle. L'issue de cette thèse, à l'heure de la généralisation de la médecine personnalisée et de l'exploitation des données massives, confirme l'importance du rôle de l'ensemble des professionnels gravitant autour du parcours de soins des enfants vulnérables. D'un point de vue de santé publique, la production des systèmes d'information de suivi d'enfants vulnérables, de périnatalité ou plus généralement de parcours patient, induit inéluctablement une coopération pluridisciplinaire de professionnels médicaux, paramédicaux, administratifs, spécialistes des questions épidémiologiques et de traitement de la donnée. Le soin du patient se trouve désormais au cœur de l'interaction entre toutes ces entités professionnelles. Cependant, au même titre que l'avis médical, le consentement et la volonté justifiée et informée du patient est désormais partie prenante dans le cadre d'une relation soignant – soigné. Les algorithmes présentés ici, au même titre que les algorithmes d'IA déjà disponibles auprès du grand public, ne sont finalement que des outils supplémentaires d'aide à la décision pour les professionnels et les patients soucieux de leur état de santé.

BIBLIOGRAPHIE

1. United Nations. Millennium Development Goals [Internet]. [cited 2023 Feb 21]. Available from: <https://www.un.org/millenniumgoals/>
2. United Nations. Reduce child mortality [Internet]. [cited 2023 Feb 21]. Available from: <https://www.un.org/en/ecosoc/about/mdg4.shtml>
3. CME Info - Child Mortality Estimates [Internet]. [cited 2023 Mar 2]. Available from: <https://childmortality.org/>
4. GBD 2015 Child Mortality Collaborators. Global, regional, national, and selected subnational levels of stillbirths, neonatal, infant, and under-5 mortality, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016 Oct 8;388(10053):1725–1774.
5. Take Action for the Sustainable Development Goals - United Nations Sustainable Development [Internet]. [cited 2023 Mar 1]. Available from: <https://www.un.org/sustainabledevelopment/sustainable-development-goals/>
6. New World Bank country classifications by income level: 2022-2023 [Internet]. [cited 2023 Jul 4]. Available from: <https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2022-2023>
7. Bremberg SG. The rate of country-level improvements of the infant mortality rate is mainly determined by previous history. *Eur J Public Health*. 2016 Aug;26(4):597–601.
8. La mortalité infantile est stable depuis dix ans après des décennies de baisse - Insee Focus - 117 [Internet]. [cited 2023 Mar 2]. Available from: <https://www.insee.fr/fr/statistiques/3560308>
9. Trinh NTH, de Visme S, Cohen JF, Bruckner T, Lelong N, Adnot P, et al. Recent historic increase of infant mortality in France: A time-series analysis, 2001 to 2019. *Lancet Reg Health Eur*. 2022 May;16:100339.

10. Liu L, Johnson HL, Cousens S, Perin J, Scott S, Lawn JE, et al. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet*. 2012 Jun 9;379(9832):2151–2161.
11. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. *Lancet*. 2008 Jan 5;371(9606):75–84.
12. Preterm birth [Internet]. [cited 2023 Mar 7]. Available from: <https://www.who.int/news-room/fact-sheets/detail/preterm-birth>
13. Blencowe H, Cousens S, Oestergaard MZ, Chou D, Moller A-B, Narwal R, et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *Lancet*. 2012 Jun 9;379(9832):2162–2172.
14. Chawanpaiboon S, Vogel JP, Moller A-B, Lumbiganon P, Petzold M, Hogan D, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob Health*. 2019 Jan;7(1):e37–e46.
15. Delnord M, Zeitlin J. Epidemiology of late preterm and early term births - An international perspective. *Semin Fetal Neonatal Med*. 2019 Feb;24(1):3–10.
16. Michel M, Alberti C, Carel J-C, Chevreul K. Social inequalities in access to care at birth and neonatal mortality: an observational study. *Arch Dis Child Fetal Neonatal Ed*. 2022 Jul;107(4):380–385.
17. Rouget F, Lebreton J, Kadhel P, Monfort C, Bodeau-Livinec F, Janky E, et al. Medical and sociodemographic risk factors for preterm birth in a French Caribbean population of African descent. *Matern Child Health J*. 2013 Aug;17(6):1103–1111.
18. Bonet M, Smith LK, Pilkington H, Draper ES, Zeitlin J. Neighbourhood deprivation and very preterm birth in an English and French cohort. *BMC Pregnancy Childbirth*. 2013 Apr 25;13:97.
19. Torchin H, Ancel PY. [Epidemiology and risk factors of preterm birth]. *J Gynecol Obstet Biol Reprod (Paris)*. 2016 Dec;45(10):1213–1230.

20. Crequit S, Chatzistergiou K, Bierry G, Bouali S, La Tour AD, Sgihouar N, et al. Association between social vulnerability profiles, prenatal care use and pregnancy outcomes. *BMC Pregnancy Childbirth*. 2023 Jun 22;23(1):465.
21. Meis PJ, Michielutte R, Peters TJ, Wells HB, Sands RE, Coles EC, et al. Factors associated with preterm birth in Cardiff, Wales. *Am J Obstet Gynecol*. 1995 Aug;173(2):597–602.
22. Chen Y, Lin H-Y, Tseng T-S, Wen H, DeVivo MJ. Racial differences in data quality and completeness: spinal cord injury model systems' experiences. *Top Spinal Cord Inj Rehabil*. 2018;24(2):110–120.
23. Blomberg M, Birch Tyrberg R, Kjølhede P. Impact of maternal age on obstetric and neonatal outcome with emphasis on primiparous adolescents and older women: a Swedish Medical Birth Register Study. *BMJ Open*. 2014 Nov 11;4(11):e005840.
24. Ruiz M, Goldblatt P, Morrison J, Kukla L, Švancara J, Riitta-Järvelin M, et al. Mother's education and the risk of preterm and small for gestational age birth: a DRIVERS meta-analysis of 12 European cohorts. *J Epidemiol Community Health*. 2015 Sep;69(9):826–833.
25. Chen X-K, Wen SW, Fleming N, Demissie K, Rhoads GG, Walker M. Teenage pregnancy and adverse birth outcomes: a large population based retrospective cohort study. *Int J Epidemiol*. 2007 Apr;36(2):368–373.
26. Thomson K, Moffat M, Arisa O, Jesurasa A, Richmond C, Odeniyi A, et al. Socioeconomic inequalities and adverse pregnancy outcomes in the UK and Republic of Ireland: a systematic review and meta-analysis. *BMJ Open*. 2021 Mar 15;11(3):e042753.
27. Santos S, Voerman E, Amiano P, Barros H, Beilin LJ, Bergström A, et al. Impact of maternal body mass index and gestational weight gain on pregnancy complications: an individual participant data meta-analysis of European, North American and Australian cohorts. *BJOG*. 2019 Jul;126(8):984–995.
28. Roman A, Ramirez A, Fox NS. Prevention of preterm birth in twin pregnancies. *American Journal of Obstetrics & Gynecology MFM*. 2022 Mar;4(2S):100551.

29. Malaza N, Masete M, Adam S, Dias S, Nyawo T, Pheiffer C. A Systematic Review to Compare Adverse Pregnancy Outcomes in Women with Pregestational Diabetes and Gestational Diabetes. *Int J Environ Res Public Health*. 2022 Aug 31;19(17).
30. Kong L, Nilsson IAK, Gissler M, Lavebratt C. Associations of maternal diabetes and body mass index with offspring birth weight and prematurity. *JAMA Pediatr*. 2019 Feb 25;173(4):371–378.
31. Battarbee AN, Sinkey RG, Harper LM, Oparil S, Tita ATN. Chronic hypertension in pregnancy. *Am J Obstet Gynecol*. 2020 Jun;222(6):532–541.
32. Tita AT, Szychowski JM, Boggess K, Dugoff L, Sibai B, Lawrence K, et al. Treatment for Mild Chronic Hypertension during Pregnancy. *N Engl J Med*. 2022 May 12;386(19):1781–1792.
33. Helmo FR, Alves EAR, Moreira RA de A, Severino VO, Rocha LP, Monteiro MLGDR, et al. Intrauterine infection, immune system and premature birth. *J Matern Fetal Neonatal Med*. 2018 May;31(9):1227–1233.
34. Hernandez-Andrade E, Maymon E, Luewan S, Bhatti G, Mehrmohammadi M, Erez O, et al. A soft cervix, categorized by shear-wave elastography, in women with short or with normal cervical length at 18-24 weeks is associated with a higher prevalence of spontaneous preterm delivery. *J Perinat Med*. 2018 Jul 26;46(5):489–501.
35. Berghella V, Bega G, Tolosa JE, Berghella M. Ultrasound assessment of the cervix. *Clin Obstet Gynecol*. 2003 Dec;46(4):947–962.
36. Adams MM, Elam-Evans LD, Wilson HG, Gilbertz DA. Rates of and factors associated with recurrence of preterm delivery. *JAMA*. 2000 Mar 29;283(12):1591–1596.
37. Kazemier BM, Buijs PE, Mignini L, Limpens J, de Groot CJM, Mol BWJ, et al. Impact of obstetric history on the risk of spontaneous preterm birth in singleton and multiple pregnancies: a systematic review. *BJOG*. 2014 Sep;121(10):1197–208; discussion 1209.
38. Iams JD, Berghella V. Care for women with prior preterm birth. *Am J Obstet Gynecol*. 2010 Aug;203(2):89–100.

39. Robison RG, Kumar R, Arguelles LM, Hong X, Wang G, Apollon S, et al. Maternal smoking during pregnancy, prematurity and recurrent wheezing in early childhood. *Pediatr Pulmonol*. 2012 Jul;47(7):666–673.
40. Bertin M, Chevrier C, Serrano T, Monfort C, Rouget F, Cordier S, et al. Association between prenatal exposure to traffic-related air pollution and preterm birth in the PELAGIE mother-child cohort, Brittany, France. Does the urban-rural context matter? *Environ Res*. 2015 Oct;142:17–24.
41. Son J-Y, Choi HM, Miranda ML, Bell ML. Exposure to heat during pregnancy and preterm birth in North Carolina: Main effect and disparities by residential greenness, urbanicity, and socioeconomic status. *Environ Res*. 2022 Mar;204(Pt C):112315.
42. Glover AV, Manuck TA. Screening for spontaneous preterm birth and resultant therapies to reduce neonatal morbidity and mortality: A review. *Semin Fetal Neonatal Med*. 2018 Apr;23(2):126–132.
43. Mathews TJ, Driscoll AK. Trends in Infant Mortality in the United States, 2005-2014. *NCHS Data Brief*. 2017 Mar;(279):1–8.
44. Aarnoudse-Moens CSH, Weisglas-Kuperus N, van Goudoever JB, Oosterlaan J. Meta-analysis of neurobehavioral outcomes in very preterm and/or very low birth weight children. *Pediatrics*. 2009 Aug;124(2):717–728.
45. Larroque B, Ancel P-Y, Marret S, Marchand L, André M, Arnaud C, et al. Neurodevelopmental disabilities and special care of 5-year-old children born before 33 weeks of gestation (the EIPAGE study): a longitudinal cohort study. *Lancet*. 2008 Mar 8;371(9615):813–820.
46. Casey PH. Growth of low birth weight preterm children. *Semin Perinatol*. 2008 Feb;32(1):20–27.
47. Cooke RWI, Foulder-Hughes L. Growth impairment in the very preterm and cognitive and motor performance at 7 years. *Arch Dis Child*. 2003 Jun;88(6):482–487.

48. Bayman E, Drake AJ, Piyasena C. Prematurity and programming of cardiovascular disease risk: a future challenge for public health? *Arch Dis Child Fetal Neonatal Ed.* 2014 Nov;99(6):F510–4.
49. Kono Y, Neonatal Research Network of Japan. Neurodevelopmental outcomes of very low birth weight infants in the Neonatal Research Network of Japan: importance of neonatal intensive care unit graduate follow-up. *Clin Exp Pediatr.* 2021 Jul;64(7):313–321.
50. Eves R, Mendonça M, Baumann N, Ni Y, Darlow BA, Horwood J, et al. Association of Very Preterm Birth or Very Low Birth Weight With Intelligence in Adulthood: An Individual Participant Data Meta-analysis. *JAMA Pediatr.* 2021 Aug 1;175(8):e211058.
51. Pascal A, Govaert P, Oostra A, Naulaers G, Ortibus E, Van den Broeck C. Neurodevelopmental outcome in very preterm and very-low-birthweight infants born over the past decade: a meta-analytic review. *Dev Med Child Neurol.* 2018 Apr;60(4):342–355.
52. Chehade H, Simeoni U, Guignard J-P, Boubred F. Preterm birth: long term cardiovascular and renal consequences. *Curr Pediatr Rev.* 2018;14(4):219–226.
53. Annesi-Maesano I, Sterlin C, Caillaud D, de Blay F, Lavaud F, Charpin D, et al. Factors related to under-diagnosis and under-treatment of childhood asthma in metropolitan France. *Multidiscip Respir Med.* 2012 Aug 8;7(1):24.
54. Franz AP, Bolat GU, Bolat H, Matijasevich A, Santos IS, Silveira RC, et al. Attention-Deficit/Hyperactivity Disorder and Very Preterm/Very Low Birth Weight: A Meta-analysis. *Pediatrics.* 2018 Jan;141(1).
55. Treyvaud K, Ure A, Doyle LW, Lee KJ, Rogers CE, Kidokoro H, et al. Psychiatric outcomes at age seven for very preterm children: rates and predictors. *J Child Psychol Psychiatry.* 2013 Jul;54(7):772–779.
56. Wheelock MD, Lean RE, Bora S, Melzer TR, Eggebrecht AT, Smyser CD, et al. Functional Connectivity Network Disruption Underlies Domain-Specific Impairments in Attention for Children Born Very Preterm. *Cereb Cortex.* 2021 Jan 5;31(2):1383–1394.

57. Walton K, McGee M, Sato J, Law N, Hopperton KE, Bando N, et al. Social-emotional functioning and dietary intake among children born with a very low birth weight. *Appl Physiol Nutr Metab*. 2022 Jul 1;47(7):737–748.
58. Treyvaud K. Parent and family outcomes following very preterm or very low birth weight birth: a review. *Semin Fetal Neonatal Med*. 2014 Apr;19(2):131–135.
59. Pace CC, Spittle AJ, Molesworth CM-L, Lee KJ, Northam EA, Cheong JLY, et al. Evolution of depression and anxiety symptoms in parents of very preterm infants during the newborn period. *JAMA Pediatr*. 2016 Sep 1;170(9):863–870.
60. Treyvaud K, Spittle A, Anderson PJ, O'Brien K. A multilayered approach is needed in the NICU to support parents after the preterm birth of their infant. *Early Hum Dev*. 2019 Dec;139:104838.
61. Singer LT, Salvator A, Guo S, Collin M, Lilien L, Baley J. Maternal psychological distress and parenting stress after the birth of a very low-birth-weight infant. *JAMA*. 1999 Mar 3;281(9):799–805.
62. Bérard A, Le Tiec M, De Vera MA. Study of the costs and morbidities of late-preterm birth. *Arch Dis Child Fetal Neonatal Ed*. 2012 Sep;97(5):F329–34.
63. Korvenranta E, Lehtonen L, Rautava L, Häkkinen U, Andersson S, Gissler M, et al. Impact of very preterm birth on health care costs at five years of age. *Pediatrics*. 2010 May;125(5):e1109–14.
64. Petrou S, Abangma G, Johnson S, Wolke D, Marlow N. Costs and health utilities associated with extremely preterm birth: evidence from the EPICure study. *Value Health*. 2009 Jul 29;12(8):1124–1134.
65. Doyle LW, Anderson PJ, Battin M, Bowen JR, Brown N, Callanan C, et al. Long term follow up of high risk children: who, why and how? *BMC Pediatr*. 2014 Nov 17;14:279.
66. L Orton J, McGinley JL, Fox LM, Spittle AJ. Challenges of neurodevelopmental follow-up for extremely preterm infants at two years. *Early Hum Dev*. 2015 Dec;91(12):689–694.
67. National Guideline Alliance (UK). Developmental follow-up of children and young people born preterm. London: National Institute for Health and Care Excellence (NICE); 2017.

68. Brett J, Staniszewska S, Newburn M, Jones N, Taylor L. A systematic mapping review of effective interventions for communicating with, supporting and providing information to parents of preterm infants. *BMJ Open*. 2011 Jun 2;1(1):e000023.
69. American Academy of Pediatrics Committee on Fetus and Newborn. Hospital discharge of the high-risk neonate. *Pediatrics*. 2008 Nov;122(5):1119–1126.
70. ASQ®:SE-2 User's Guide [Internet]. [cited 2023 Mar 17]. Available from: <https://products.brookespublishing.com/ASQSE-2-Users-Guide-P851.aspx>
71. Atger-Lallier L, Guilmin-Crepon S, Boizeau P, Zenaty D, Simon D, Paulsen A, et al. Factors Affecting Loss to Follow-Up in Children and Adolescents with Chronic Endocrine Conditions. *Horm Res Paediatr*. 2019;92(4):254–261.
72. Swearingen C, Simpson P, Cabacungan E, Cohen S. Social disparities negatively impact neonatal follow-up clinic attendance of premature infants discharged from the neonatal intensive care unit. *J Perinatol*. 2020 May;40(5):790–797.
73. Mas C, Gérardin P, Chirpaz E, Carbonnier M, Mussard C, Samperiz S, et al. Follow-up at two years of age and early predictors of non-compliance in a cohort of very preterm infants. *Early Hum Dev*. 2017 Mar 22;108:1–7.
74. Guillén U, DeMauro S, Ma L, Zupancic J, Roberts R, Schmidt B, et al. Relationship between attrition and neurodevelopmental impairment rates in extremely preterm infants at 18 to 24 months: a systematic review. *Arch Pediatr Adolesc Med*. 2012 Feb;166(2):178–184.
75. Teixeira R, Queiroga AC, Freitas AI, Lorthe E, Santos AC, Moreira C, et al. Completeness of retention data and determinants of attrition in birth cohorts of very preterm infants: A systematic review. *Front Pediatr*. 2021 Feb 17;9:529733.
76. Hanf M, Nusinovici S, Rouger V, Olivier M, Berlie I, Flamant C, et al. Cohort Profile: Longitudinal study of preterm infants in the Pays de la Loire region of France (LIFT cohort). *Int J Epidemiol*. 2017 Oct 1;46(5):1396–1397h.
77. Pierrat V, Marchand-Martin L, Marret S, Arnaud C, Benhammou V, Cambonie G, et al. Neurodevelopmental outcomes at age 5 among children born preterm: EPIPAGE-2 cohort study. *BMJ*. 2021 Apr 28;373:n741.

78. Greene N, Greenland S, Olsen J, Nohr EA. Estimating bias from loss to follow-up in the Danish National Birth Cohort. *Epidemiology*. 2011 Nov;22(6):815–822.
79. Hintz SR, Gould JB, Bennett MV, Lu T, Gray EE, Jocson MAL, et al. Factors Associated with Successful First High-Risk Infant Clinic Visit for Very Low Birth Weight Infants in California. *J Pediatr*. 2019 Apr 6;210:91–98.e1.
80. Shin SH, Sohn JA, Kim E-K, Shin SH, Kim H-S, Lee JA. Factors associated with the follow-up of high risk infants discharged from a neonatal intensive care unit. *Pediatr Neonatol*. 2022 Jul;63(4):373–379.
81. La médecine : vision d'aujourd'hui et de demain | Leem [Internet]. [cited 2023 Sep 20]. Available from: <https://www.leem.org/la-medecine-vision-daujourdhui-et-de-demain#:~:text=La%20sant%C3%A9%20devient%20de%20plus,d%C3%A9veloppeur%20telle%20ou%20telle%20maladie>
82. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015 Feb 26;372(9):793–795.
83. Khoury MJ, Lademaro MF, Riley WT. Precision public health for the era of precision medicine. *Am J Prev Med*. 2016 Mar;50(3):398–401.
84. Rubin R. Precision medicine: the future or simply politics? *JAMA*. 2015 Mar 17;313(11):1089–1091.
85. König IR, Fuchs O, Hansen G, von Mutius E, Kopp MV. What is precision medicine? *Eur Respir J*. 2017 Oct 19;50(4).
86. Khoury MJ. Planning for the future of epidemiology in the era of big data and precision medicine. *Am J Epidemiol*. 2015 Dec 15;182(12):977–979.
87. Goetz LH, Schork NJ. Personalized medicine: motivation, challenges, and progress. *Fertil Steril*. 2018 Jun;109(6):952–963.
88. Kasztura M, Richard A, Bempong N-E, Loncar D, Flahault A. Cost-effectiveness of precision medicine: a scoping review. *Int J Public Health*. 2019 Dec;64(9):1261–1271.

89. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol*. 2011 Mar;8(3):184–187.
90. Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *N Biotechnol*. 2012 Sep 15;29(6):613–624.
91. Hood L. Systems biology and p4 medicine: past, present, and future. *Rambam Maimonides Med J*. 2013 Apr 30;4(2):e0012.
92. Pravettoni G, Gorini A. A P5 cancer medicine approach: why personalized medicine cannot ignore psychology. *J Eval Clin Pract*. 2011 Aug 1;17(4):594–596.
93. Gorini A, Pravettoni G. P5 medicine: a plus for a personalized approach to oncology. *Nat Rev Clin Oncol*. 2011 Jul;8(7):444–444.
94. Ibrahim MS, Mohamed Yusoff H, Abu Bakar YI, Thwe Aung MM, Abas MI, Ramli RA. Digital health for quality healthcare: A systematic mapping of review studies. *Digit Health*. 2022 Mar 18;8:20552076221085810.
95. La médecine de demain : digitalisation des parcours de soin et symbiose de l'ensemble des acteurs de la santé - Salesforce Blog France [Internet]. [cited 2023 Sep 20]. Available from: <https://www.salesforce.com/fr/blog/2022/03/La-medecine-de-demain-Digitalisation-des-parcours-de-soin-et-symbiose-de-l-ensemble-acteurs-de-sante.html>
96. The Six P's Of Medicine: How Healthcare Technology Will Transform Access To Care And Enable Better Patient Outcomes [Internet]. [cited 2023 Sep 20]. Available from: <https://www.forbes.com/sites/forbestechcouncil/2021/03/30/the-six-ps-of-medicine-how-healthcare-technology-will-transform-access-to-care-and-enable-better-patient-outcomes/?sh=14cf82bd2ebe>
97. An Early Information Society: News and the Media in Eighteenth-Century Paris. *Am Hist Rev*. 2000 Feb;
98. Data growth worldwide 2010-2025 | Statista [Internet]. [cited 2023 Sep 20]. Available from: <https://www.statista.com/statistics/871513/worldwide-data-created/>

99. Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc*. 2014 Dec;21(6):957–958.
100. Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. *Biomed Res Int*. 2015 Jun 2;2015:639021.
101. Favaretto M, De Clercq E, Schneble CO, Elger BS. What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLoS One*. 2020 Feb 25;15(2):e0228987.
102. Wong HT, Chiang VCL, Choi KS, Loke AY. The need for a definition of big data for nursing science: A case study of disaster preparedness. *Int J Environ Res Public Health*. 2016 Oct 17;13(10).
103. 3D Data Management: Controlling Data Volume, Velocity, and Variety [Internet]. [cited 2023 Sep 20]. Available from: <https://studylib.net/doc/8647594/3d-data-management--controlling-data-volume--velocity--an...>
104. Zhang Y, Zhao Y. Astronomy in the big data era. *Data Sci J*. 2015 May 22;14(0):11.
105. Goldstein I, Spatt CS, Ye M. Big data in finance. *Rev Financ Stud*. 2021 Apr 5;
106. Cockcroft S, Russell M. Big data opportunities for accounting and finance practice and research. *Australian Accounting Review*. 2018 Sep;28(3):323–333.
107. How Big Data Is Helping Advertisers Solve Problems [Internet]. [cited 2023 Sep 22]. Available from: <https://www.forbes.com/sites/forbesagencycouncil/2022/03/15/how-big-data-is-helping-advertisers-solve-problems/?sh=2ca79e3625e0>
108. Saidali J, Rahich H, Tabaa Y, Medouri A. The combination between Big Data and Marketing Strategies to gain valuable Business Insights for better Production Success. *Procedia Manufacturing*. 2019;32:1017–1023.
109. Zhang J, Wolfram D, Ma F. The impact of big data on research methods in information science. *Data Inf Manag*. 2023 Jun;7(2):100038.

110. O'Donovan P, Leahy K, Bruton K, O'Sullivan DTJ. Big data in manufacturing: a systematic mapping study. *J Big Data*. 2015 Dec;2(1):20.
111. van der Voort HG (Haiko), Klievink AJ (Bram), Arnaboldi M (Michela), Meijer AJ (Albert). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Gov Inf Q*. 2018 Oct;
112. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. *J Big Data*. 2015 Dec;2(1):1.
113. Ginsburg GS, Phillips KA. Precision medicine: from science to value. *Health Aff (Millwood)*. 2018 May;37(5):694–701.
114. Flott K, Darzi A, Mayer E. Care pathway and organisational features driving patient experience: statistical analysis of large NHS datasets. *BMJ Open*. 2018 Jul 7;8(7):e020411.
115. Moreira A, Tovar M, Smith AM, Lee GC, Meunier JA, Cheema Z, et al. Development of a peripheral blood transcriptomic gene signature to predict bronchopulmonary dysplasia. *Am J Physiol Lung Cell Mol Physiol*. 2023 Jan 1;324(1):L76–L87.
116. Pammi M, Aghaeepour N, Neu J. Multiomics, artificial intelligence, and precision medicine in perinatology. *Pediatr Res*. 2023 Jan;93(2):308–315.
117. Mussap M, Ferrari M, Fanos V. Laboratory medicine meets precision medicine: The paradigm of metabolomics in perinatology: Proposals from the 4th International Conference on Neonatal and Pediatric Laboratory Medicine. *Clin Chim Acta*. 2015 Dec 7;451(Pt A):1–3.
118. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018 Oct 10;2(10):719–731.
119. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol*. 2019 May;20(5):e262–e273.
120. Deo RC. Machine learning in medicine. *Circulation*. 2015 Nov 17;132(20):1920–1930.
121. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018 Apr 3;319(13):1317–1318.

122. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol*. 2019 Dec 31;188(12):2222–2239.
123. Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open*. 2020 Mar 2;3(3):e200265.
124. Bora A, Balasubramanian S, Babenko B, Virmani S, Venugopalan S, Mitani A, et al. Predicting the risk of developing diabetic retinopathy using deep learning. *Lancet Digit Health*. 2021 Jan;3(1):e10–e19.
125. Wu K, Wu E, DAndrea M, Chitale N, Lim M, Dabrowski M, et al. Machine learning prediction of clinical trial operational efficiency. *AAPS J*. 2022 Apr 21;24(3):57.
126. Moore PJ, Lyons TJ, Gallacher J, Alzheimer's Disease Neuroimaging Initiative. Random forest prediction of Alzheimer's disease using pairwise selection from time series data. *PLoS One*. 2019 Feb 14;14(2):e0211558.
127. Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. *Nat Rev Urol*. 2019 Jul;16(7):391–403.
128. Lloyd S. Least squares quantization in PCM. *IEEE Trans Inform Theory*. 1982 Mar;28(2):129–137.
129. Hand DJ, McLachlan GJ, Basford KE. Mixture models: inference and applications to clustering. *Appl Stat*. 1989;38(2):384.
130. Nugent R, Meila M. An overview of clustering applied to molecular biology. *Methods Mol Biol*. 2010;620:369–404.
131. Abdi H, Williams LJ. Principal component analysis. *WIREs Comp Stat*. 2010 Jul;2(4):433–459.
132. Martino DJ, Valerio MP, Parker G. The structure of mania: An overview of factorial analysis studies. *Eur Psychiatry*. 2020 Feb 10;63(1):e10.
133. Chowdhury GG. Natural language processing. *Ann Rev Info Sci Tech*. 2005 Jan 31;37(1):51–89.

134. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. Wiley; 2013.
135. Eliason SR. Maximum Likelihood Estimation: Logic and Practice . 1993.
136. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996 Jan;58(1):267–288.
137. Friedman JH. Multivariate adaptive regression splines. Ann Statist. 1991 Mar;19(1):1–67.
138. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.
139. Breiman L. Bagging predictors. Mach Learn. 1996 Aug;24(2):123–140.
140. Breiman L. Random Forests. Springer Science and Business Media LLC. 2001;
141. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'16. New York, New York, USA: ACM Press; 2016. p. 785–794.
142. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist. 2001 Oct;29(5):1189–1232.
143. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep;20(3):273–297.
144. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943 Dec;5(4):115–133.
145. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–536.
146. Ruder S. An overview of gradient descent optimization algorithms. arXiv. 2016;
147. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014 Jan;15(1):1929–1958.
148. Wolpert DH. Stacked generalization. Neural Netw. 1992 Jan;5(2):241–259.

149. Whalen S, Pandey OP, Pandey G. Predicting protein function and other biomedical characteristics with heterogeneous ensembles. *Methods*. 2016 Jan 15;93:92–102.
150. Yan Y, Schaffter T, Bergquist T, Yu T, Prosser J, Aydin Z, et al. A Continuously Benchmarked and Crowdsourced Challenge for Rapid Development and Evaluation of Models to Predict COVID-19 Diagnosis and Hospitalization. *JAMA Netw Open*. 2021 Oct 1;4(10):e2124946.
151. Hsu LL, Culhane AC. Impact of data preprocessing on integrative matrix factorization of single cell data. *Front Oncol*. 2020 Jun 23;10:973.
152. Rubin DB, editor. *Multiple imputation for nonresponse in surveys*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 1987.
153. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inform Theory*. 1967 Jan;13(1):21–27.
154. Box GEP, Cox DR. *An Analysis of Transformations*. JSTOR. 1964;26.
155. Nematzadeh S, Kiani F, Torkamanian-Afshar M, Aydin N. Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases. *Comput Biol Chem*. 2022 Apr;97:107619.
156. Jung Y, Hu J. A K-fold Averaging Cross-validation Procedure. *J Nonparametr Stat*. 2015 Feb 26;27(2):167–179.
157. Janssens ACJW, Martens FK. Reflection on modern methods: Revisiting the area under the ROC Curve. *Int J Epidemiol*. 2020 Aug 1;49(4):1397–1403.
158. Accueil | Les services de l'État en région [Internet]. [cited 2023 Apr 19]. Available from: <https://www.prefectures-regions.gouv.fr/>
159. INSTRUCTION N°DGOS/PF3/R3/DGS/MC1/2015/227 du 3 juillet 2015 relative à l'actualisation et à l'harmonisation des missions des réseaux de santé en périnatalité dans un cadre régional. - Légifrance [Internet]. [cited 2023 Apr 19]. Available from: <https://www.legifrance.gouv.fr/circulaire/id/39846>

160. FFRSP – Fédération Française des Réseaux de Santé en Périnatalité [Internet]. [cited 2023 Apr 24]. Available from: <https://ffrsp.fr/>
161. Tallec A, Baudier F, Czernichow P, Michaud C, Pomarède R. Les politiques régionales de santé publique: Éléments de constat et préconisations dans le contexte de la loi hôpital-patients-santé-territoires. HCSP. 2008 Oct 2;
162. Seppänen A-V, Draper ES, Petrou S, Barros H, Andronis L, Kim SW, et al. Follow-up after very preterm birth in Europe. *Arch Dis Child Fetal Neonatal Ed.* 2022 Jan;107(1):113–114.
163. van Boven MR, Henke CE, Leemhuis AG, Hoogendoorn M, van Kaam AH, Königs M, et al. Machine learning prediction models for neurodevelopmental outcome after preterm birth: A scoping review and new machine learning evaluation framework. *Pediatrics.* 2022 Jul 1;150(1).
164. Sesan - Service de santé numérique [Internet]. [cited 2023 Jul 24]. Available from: <https://www.sesan.fr/>
165. Base des unités urbaines 2020 | Insee [Internet]. [cited 2022 Mar 16]. Available from: <https://www.insee.fr/fr/information/4802589>
166. Fenton TR, Nasser R, Eliasziw M, Kim JH, Bilan D, Sauve R. Validating the weight gain of preterm infants between the reference growth curve of the fetus and the term infant. *BMC Pediatr.* 2013 Jun 11;13:92.
167. pcs2003-1-Agriculteurs exploitants | Insee [Internet]. [cited 2023 Jul 4]. Available from: <https://www.insee.fr/fr/metadonnees/pcs2003/categorieSocioprofessionnelleAgreee/1?champRecherche=true>
168. Flamant C, Branger B, Nguyen The Tich S, de la Rochebrochard E, Savagner C, Berlie I, et al. Parent-completed developmental screening in premature children: a valid tool for follow-up programs. *PLoS One.* 2011 May 26;6(5):e20004.
169. En 2017, l'Île-de-France reste la région où les écarts de niveaux de vie sont les plus importants, notamment à Paris - Insee Analyses Ile-de-France - 135 [Internet]. [cited 2023 May 1]. Available from: <https://www.insee.fr/fr/statistiques/5369176>

170. Déménager pour une autre région - Insee Première - 1540 [Internet]. [cited 2023 May 1]. Available from: <https://www.insee.fr/fr/statistiques/1288041>
171. Bilan démographique 2022 de l'Île-de-France : deux naissances pour un décès - Insee Flash Ile-de-France - 77 [Internet]. [cited 2023 Jul 6]. Available from: <https://www.insee.fr/fr/statistiques/6968304>
172. Groupe de travail Perinat-ARS-IDF. Rapport 2017 Périnat-ARS-IDF.
173. Müller J-B, Hanf M, Flamant C, Olivier M, Rouger V, Gascoin G, et al. Relative contributions of prenatal complications, perinatal characteristics, neonatal morbidities and socio-economic conditions of preterm infants on the occurrence of developmental disorders up to 7 years of age. *Int J Epidemiol*. 2019 Feb 1;48(1):71–82.
174. Linsell L, Malouf R, Morris J, Kurinczuk JJ, Marlow N. Prognostic factors for poor cognitive development in children born very preterm or with very low birth weight: A systematic review. *JAMA Pediatr*. 2015 Dec;169(12):1162–1172.
175. Pizer J, ElBassiouny A. Wechsler preschool and primary scale of intelligence (WPPSI). In: Carducci BJ, Nave CS, Nave CS, editors. *The wiley encyclopedia of personality and individual differences: models and theories*. Wiley; 2020. p. 473–475.
176. Bayley N. *Bayley Scales of Infant and Toddler Development, Third Edition*. American Psychological Association (APA). 2012;
177. Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. *Rev Epidemiol Sante Publique*. 2017 Oct;65 Suppl 4:S149–S167.
178. Kim NH, Youn YA, Cho SJ, Hwang J-H, Kim E-K, Kim EA-R, et al. The predictors for the non-compliance to follow-up among very low birth weight infants in the Korean neonatal network. *PLoS One*. 2018 Oct 1;13(10):e0204421.

179. Patra K, Greene M, Pérez B, Silvestri J. Neonatal High-Risk Follow-up Clinics: How to Improve Attendance in Very Low Birth Weight Infants. *E-Journal of Neonatology Research*. 2014;4(1):3–13.
180. MacBean V, Drysdale SB, Zivanovic S, Peacock JL, Greenough A. Participant retention in follow-up studies of prematurely born children. *BMC Public Health*. 2019 Sep 6;19(1):1233.
181. DeMauro SB, Bellamy SL, Fernando M, Hoffmann J, Gratton T, Schmidt B, et al. Patient, Family, and Center-Based Factors Associated with Attrition in Neonatal Clinical Research: A Prospective Study. *Neonatology*. 2019 Mar 5;115(4):328–334.
182. Fenton TR, Anderson D, Groh-Wargo S, Hoyos A, Ehrenkranz RA, Senterre T. An Attempt to Standardize the Calculation of Growth Velocity of Preterm Infants-Evaluation of Practical Bedside Methods. *J Pediatr*. 2018 May;196:77–83.
183. GitHub - amices/mice: Multivariate Imputation by Chained Equations [Internet]. [cited 2023 Oct 1]. Available from: <https://github.com/amices/mice>
184. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale)*. 2016 Apr;6(2).
185. O'Brien RM. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Qual Quant*. 2007 Sep 11;41(5):673–690.
186. Bruneau L, Billaud E, Raffi F, Hanf M. Factors associated with the level of CD4 cell counts at HIV diagnosis in a French cohort: a quantile regression approach. *Int J STD AIDS*. 2017;28(4):397–403.
187. Pierrat V, Marchand-Martin L, Arnaud C, Kaminski M, Resche-Rigon M, Lebeaux C, et al. Neurodevelopmental outcome at 2 years for preterm children born at 22 to 34 weeks' gestation in France in 2011: EPIPAGE-2 cohort study. *BMJ*. 2017 Aug 16;358:j3448.
188. Moore T, Hennessy EM, Myles J, Johnson SJ, Draper ES, Costeloe KL, et al. Neurological and developmental outcome in extremely preterm children born in England in 1995 and 2006: the EPICure studies. *BMJ*. 2012 Dec 4;345:e7961.

189. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med Res Methodol*. 2015 Apr 7;15(1):30.
190. Beckung E, Carlsson G, Carlsdotter S, Uvebrant P. The natural history of gross motor development in children with cerebral palsy aged 1 to 15 years. *Dev Med Child Neurol*. 2007 Oct;49(10):751–756.
191. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019 Apr 4;380(14):1347–1358.
192. Bertholet N, Cunningham JA. Information technology and addiction science: promises and challenges. *Addict Sci Clin Pract*. 2021 Jan 26;16(1):7.
193. Arigo D, Jake-Schoffman DE, Wolin K, Beckjord E, Hekler EB, Pagoto SL. The history and future of digital health in the field of behavioral medicine. *J Behav Med*. 2019 Mar 1;42(1):67–83.
194. Cunningham JA, Gulliver A, Farrer L, Bennett K, Carron-Arthur B. Internet interventions for mental health and addictions: current findings and future directions. *Curr Psychiatry Rep*. 2014 Dec;16(12):521.
195. Michie S, Yardley L, West R, Patrick K, Greaves F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. *J Med Internet Res*. 2017 Jun 29;19(6):e232.
196. Naylor CD. On the prospects for a (deep) learning health care system. *JAMA*. 2018 Sep 18;320(11):1099–1100.
197. Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *JAMA*. 2016 Dec 13;316(22):2353–2354.
198. Jiang P, Sinha S, Aldape K, Hannenhalli S, Sahinalp C, Ruppin E. Big data in basic and translational cancer research. *Nat Rev Cancer*. 2022 Nov;22(11):625–639.
199. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol Adv*. 2021 Mar 29;49:107739.

200. Artetxe A, Beristain A, Graña M. Predictive models for hospital readmission risk: A systematic review of methods. *Comput Methods Programs Biomed.* 2018 Oct;164:49–64.
201. Oke I, VanderVeen D. Machine learning applications in pediatric ophthalmology. *Semin Ophthalmol.* 2021 Feb 27;1–7.
202. Clarke SL, Parmesar K, Saleem MA, Ramanan AV. Future of machine learning in paediatrics. *Arch Dis Child.* 2022 Mar;107(3):223–228.
203. Reed RA, Morgan AS, Zeitlin J, Jarreau P-H, Torchin H, Pierrat V, et al. Machine-Learning vs. Expert-Opinion Driven Logistic Regression Modelling for Predicting 30-Day Unplanned Rehospitalisation in Preterm Babies: A Prospective, Population-Based Study (EPIPAGE 2). *Front Pediatr.* 2020;8:585868.
204. Masino AJ, Harris MC, Forsyth D, Ostapenko S, Srinivasan L, Bonafide CP, et al. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PLoS One.* 2019 Feb 22;14(2):e0212665.
205. Teixeira R, Rodrigues C, Moreira C, Barros H, Camacho R. Machine learning methods to predict attrition in a population-based cohort of very preterm infants. *Sci Rep.* 2022 Jun 22;12(1):10587.
206. Tidymodels - Tidymodels packages [Internet]. [cited 2023 May 8]. Available from: <https://www.tidymodels.org/packages/>
207. Bania RK, Halder A. R-Ensembler: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data. *Comput Methods Programs Biomed.* 2020 Feb;184:105122.
208. Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence.* 2003 May;17(5-6):519–533.
209. Bania RK, Halder A. R-HEFS: Rough set based heterogeneous ensemble feature selection method for medical data classification. *Artif Intell Med.* 2021 Apr;114:102049.
210. Haibo He, Yang Bai, Garcia EA, Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE; 2008. p. 1322–1328.

211. McKay MD, Beckman RJ, Conover WJ. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*. 1979 May;21(2):239.
212. Bey R, Goussault R, Grolleau F, Benchoufi M, Porcher R. Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. *J Am Med Inform Assoc*. 2020 Aug 1;27(8):1244–1251.
213. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006 Jun;27(8):861–874.
214. Tseng Y-J, Huang C-E, Wen C-N, Lai P-Y, Wu M-H, Sun Y-C, et al. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *Int J Med Inform*. 2019 Aug;128:79–86.
215. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*. 2019 Dec 21;19(1):281.
216. Wijffels J. Schedule R Scripts and Processes with the “cron” Job Scheduler [R package cronR version 0.6.5]. 2023 Jan 9;
217. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013 Feb 5;10(2):e1001381.
218. Circulaire DHOS/O1/O3/CNAMTS n° 2006-151 du 30 mars 2006 relative au cahier des charges national des réseaux de santé en périnatalité - APHP DAJDP [Internet]. [cited 2023 Apr 19]. Available from: <http://affairesjuridiques.aphp.fr/textes/circulaire-dhoso1o3cnamts-n-2006-151-du-30-mars-2006-relative-au-cahier-des-charges-national-des-reseaux-de-sante-en-perinatalite/>
219. Accouchement en Ile-de-France: lancement d’une cellule régionale pour l’accompagnement des femmes enceintes sans inscription en maternité | Agence régionale de santé Ile-de-France [Internet]. [cited 2023 Sep 29]. Available from: <https://www.iledefrance.ars.sante.fr/accouchement-en-ile-de-france-lancement-dune-cellule-regionale-pour-laccompagnement-des-femmes>

220. Glauzy A, Baertschi B, Duclos-Vallée J-C. [Medical consent in the era of personalized medicine: Issues and recommendations]. *Med Sci (Paris)*. 2023 Sep 11;39(8-9):658–663.
221. Accueil | SNDS [Internet]. [cited 2023 Jul 9]. Available from: <https://www.snds.gouv.fr/SNDS/Accueil>
222. Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv*. 2012;
223. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*. 2018;
224. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278–2324.
225. Web Application Framework for R [R package shiny version 1.7.4] [Internet]. 2022 [cited 2023 Jun 13]. Available from: <https://cran.r-project.org/web/packages/shiny/index.html>
226. Wood NS, Marlow N, Costeloe K, Gibson AT, Wilkinson AR. Neurologic and developmental disability after extremely preterm birth. EPICure Study Group. *N Engl J Med*. 2000 Aug 10;343(6):378–384.
227. Costeloe K, Hennessy E, Gibson AT, Marlow N, Wilkinson AR. The EPICure study: outcomes to discharge from hospital for infants born at the threshold of viability. *Pediatrics*. 2000 Oct;106(4):659–671.
228. Larroque B, Bréart G, Kaminski M, Dehan M, André M, Burguet A, et al. Survival of very preterm infants: Epipage, a population based cohort study. *Arch Dis Child Fetal Neonatal Ed*. 2004 Mar;89(2):F139–44.
229. Vanhaesebrouck P, Allegaert K, Bottu J, Debauche C, Devlieger H, Docx M, et al. The EPIBEL study: outcomes to discharge from hospital for extremely preterm infants in Belgium. *Pediatrics*. 2004 Sep;114(3):663–675.
230. Markestad T, Kaaresen PI, Rønnestad A, Reigstad H, Lossius K, Medbø S, et al. Early death, morbidity, and need of treatment among extremely premature infants. *Pediatrics*. 2005 May;115(5):1289–1298.

231. Leviton A, Allred EN, Kuban KCK, Hecht JL, Onderdonk AB, O'shea TM, et al. Microbiologic and histologic characteristics of the extremely preterm infant's placenta predict white matter damage and later cerebral palsy. the ELGAN study. *Pediatr Res*. 2010 Jan;67(1):95–101.
232. Kuban K, Adler I, Allred EN, Batton D, Bezinque S, Betz BW, et al. Observer variability assessing US scans of the preterm brain: the ELGAN study. *Pediatr Radiol*. 2007 Dec;37(12):1201–1208.
233. Draper ES, Zeitlin J, Fenton AC, Weber T, Gerrits J, Martens G, et al. Investigating the variations in survival rates for very preterm infants in 10 European regions: the MOSAIC birth cohort. *Arch Dis Child Fetal Neonatal Ed*. 2009 May;94(3):F158–63.
234. EXPRESS Group, Fellman V, Hellström-Westas L, Norman M, Westgren M, Källén K, et al. One-year survival of extremely preterm infants after active perinatal care in Sweden. *JAMA*. 2009 Jun 3;301(21):2225–2233.
235. Costeloe KL, Hennessy EM, Haider S, Stacey F, Marlow N, Draper ES. Short term outcomes after extreme preterm birth in England: comparison of two birth cohorts in 1995 and 2006 (the EPICure studies). *BMJ*. 2012 Dec 4;345:e7976.
236. Zeitlin J, Manktelow BN, Piedvache A, Cuttini M, Boyle E, van Heijst A, et al. Use of evidence based practices to improve survival without severe morbidity for very preterm infants: results from the EPICE population based cohort. *BMJ*. 2016 Jul 5;354:i2976.
237. Ancel P-Y, Goffinet F, EPIPAGE 2 Writing Group. EPIPAGE 2: a preterm birth cohort in France in 2011. *BMC Pediatr*. 2014 Apr 9;14:97.

ANNEXES

Annexe 1 : Tableau de synthèse des études publiées sur les enfants vulnérables

Tableau S1 : Synthèse des principales cohortes d'enfants vulnérables dans le monde

Nom de la cohorte (référence)	Nombre d'enfants inclus	Pays	Dates d'inclusion	Critères d'inclusion
EPICURE (226,227)	4 004	UK-Irlande	1995	< 26 SA
EPIPAGE (45,228)	2 573	France	1997 à 2011	22-32 SA
EPIBEL (229)	525	Belgique	1999 à 2000	< 26 SA
MBRN (230)	636	Norvège	1999 à 2000	22-27 SA
ELGAN (231,232)	1 506	USA	2002 à 2004	< 28 SA
MOSAIC (233)	7 222	10 pays européens	2003	22-31 SA
LIFT (76)	8 613	France	Depuis mars 2003	22-34 SA
EXPRESS (234)	707	Suède	2004 à 2007	< 27 SA
EPICURE-2 (235)	3 133	UK-Irlande	2006	< 26 SA
EPICE (236)	7 336	11 pays européens	2011 à 2012	24-32 SA
EPIPAGE-2 (187,237)	8 400	France	Depuis mars 2011	22-34 SA

SA : semaines d'aménorrhée

Annexe 2 : Extraction du manuel d'exploitation des données de HYGIE-SEV fourni par l'industriel

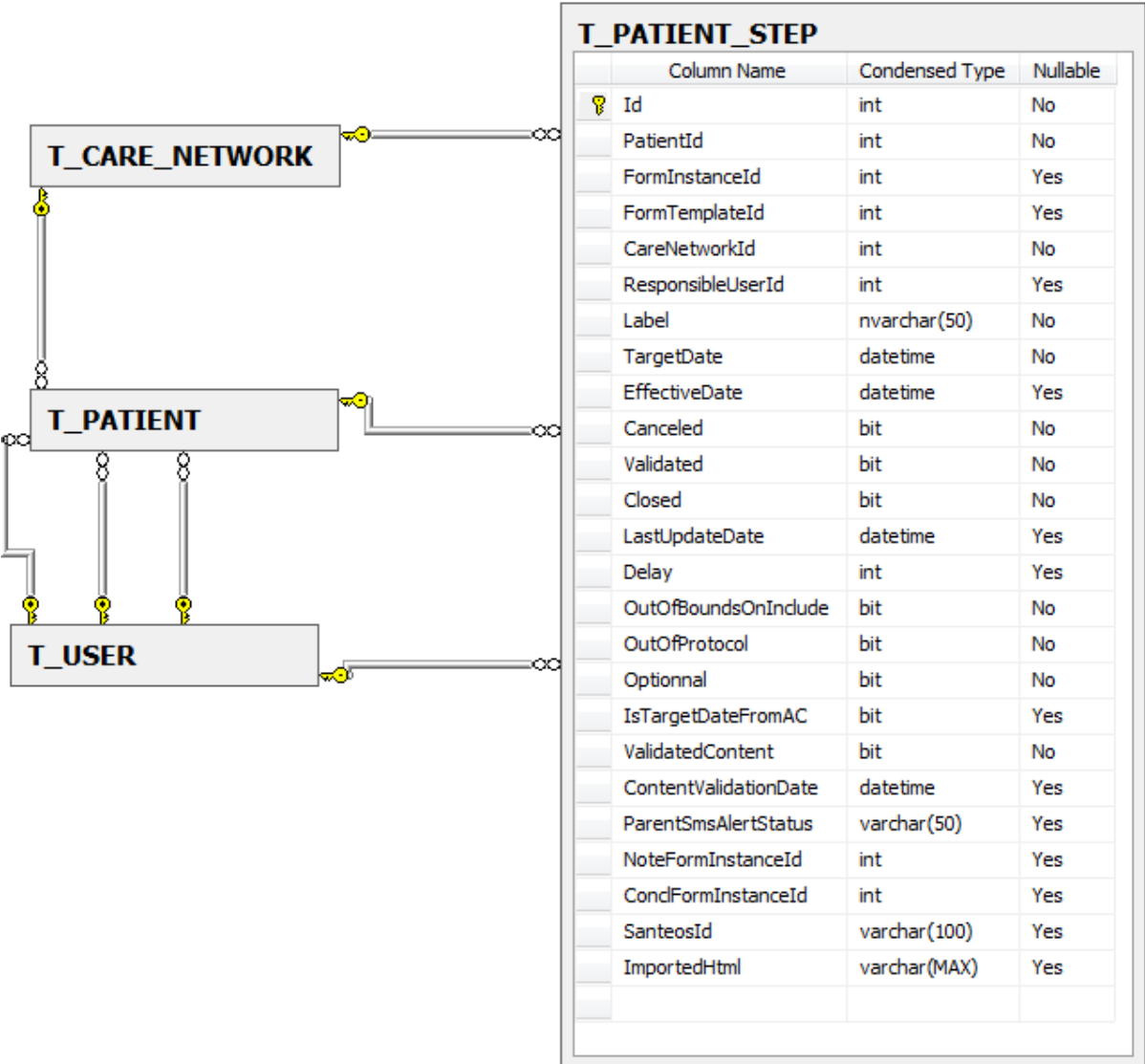


Figure S1 : Exemple de schéma de données du manuel d'exploitation fourni par l'industriel pour comprendre l'architecture de données de HYGIE-SEV

Annexe 3 : Scores utilisés dans le programme HYGIE-SEV

Tableau S2 : Description des scores à 4 mois

Type de score	Valeur du score	Libellé
Score neuro-moteur	1	Gesticulation riche, symétrique Tonus : Mis en position ventrale, redresse sa tête à 90°, prend appui sur avant-bras / Tenue de tête parfaite, angles pop 90-120° ; Adduction 70-110. Bonnes réponses antigravitaires Bonne participation aux enchainements moteurs Ouvre les mains en regardant l'objet, ébauche préhension palmaire (au contact)
	2	Bonnes acquisitions posturales mais asymétrie de tonus (fléchisseur/extenseur) ou asymétrie des angles. Motricité pauvre Réponses antigravitaires incomplètes lentes ou saccadées
	3	Hypotonie axiale, pas de redressement Hypertonie franche des membres avec retentissement sur les acquisitions posturales Motricité figée
Score relationnel	1	Répond aux sollicitations et sollicite activement Bon contact oculaire : regarde les objets et les visages autour de lui Sourire réponse Si agité, se laisse facilement apaiser
	2	Répond aux sollicitations mais ne sollicite pas S'intéresse aux objets et aux visages sur sollicitation (mais non spontanément) Si agité, long à apaiser
	3	Répond peu ou pas aux sollicitations, pas d'interactivité Absence de sourire réponse Irritabilité, difficile à gérer (inconsolable)

Tableau S3 : Description des scores à 1 an

Type de score	Valeur	Libellé
Score respiratoire	1	Aucun problème
	2	Bronchites / bronchiolites occasionnelles peu sévères (maxi 2)
	3	Bronchites / bronchiolites répétées (= 3) et/ou sévères ou wheezing fréquent
	4	Oxygénothérapie ou trachéotomie / hospitalisations multiples pour problèmes respiratoires
Autres problèmes de santé	1	Aucun problème
	2	Problèmes occasionnels facilement résolus
	3	Problèmes répétitifs ou non encore résolus
	4	Problèmes gênant la vie quotidienne ou nécessitant plusieurs hospitalisations
Score moteur	1	Se met assis seul Assis bien stable avec de bonnes réactions parachute Se met debout en passant bien par le chevalier servant Marche en appui ou seul, déplacements aisés et rapides (ex. quatre pattes) Oriente correctement la main pour une approche adaptée dans l'espace et directe de l'objet
	2	Ne se met pas assis seul mais mis assis : tient bien A trouvé un mode de déplacement ; ébauche la position petit lapin Petit retard des acquisitions mais pas d'anomalie du tonus à l'examen Approche moins adaptée de l'objet mais bon délié des doigts
	3	Mis assis : tient un moment, cyphose – réactions parachute insuffisantes Déplacements limités ou difficiles : retournements, ramper et/ou hypertonie distale des membres inférieurs ou déficit unilatéral, et/ou préhension enraidie, mal calibrée Mauvais délié des doigts
	4	Grande hypotonie axiale Pas de déplacement : ne rampe pas Diplégie sévère ou quadriplégie manifeste Préhension difficile, absence de délié des doigts
Score cognitif	1	Pince pouce index Explore attentivement les objets et porte peu à la bouche Recherche le jouet caché ostensiblement Commence à remettre la pastille dans le bocal ou les cubes dans la boîte Comprend les ordres, les interdictions Jargonne Pointe du doigt

Tableau S3 (suite)

Type de score	Valeur	Libellé
Score cognitif	2	<p>Activité exploratrice moins élaborée, porte encore beaucoup à la bouche</p> <p>Vide les boîtes et jette</p> <p>Dévisage l'étranger</p> <p>Fait bravo et/ou au revoir</p> <p>Signe d'anticipation (signe du manteau)</p> <p>Double les monosyllabes</p>
	3	<p>Peu d'explorations (porte à la bouche et jette)</p> <p>Ne différencie pas l'étranger</p> <p>Pas d'anticipation des situations ; compréhension limitée</p>
	4	<p>Manipulations absentes ou stéréotypées</p> <p>Contact médiocre à mauvais</p> <p>Indifférent à l'environnement</p>
Score relationnel	1	<p>Répond facilement aux sollicitations d'interaction et d'attention conjointe (obligatoire pour le score 1)</p> <p>Maintient l'interaction (obligatoire pour le score 1)</p> <p>Initie lui-même l'interaction</p> <p>Répond à son prénom</p> <p>Communication non verbale riche</p> <p>Mimique riche, adressée</p> <p>Imitations facilement observées</p>
	2	<p>Répond aux sollicitations mais ne maintient pas l'interaction, poursuit le jeu en solo (obligatoire pour le score 2)</p> <p>Initie peu l'interaction (obligatoire pour le score 2)</p> <p>Comportement plus « indépendant » - ne cherche pas trop le contact physique</p> <p>Communication non verbale pauvre ; mimique pauvre mais adressée</p> <p>Peu d'imitations</p> <p>Bébé trop calme et trop sérieux ou au contraire agité</p>
	3	<p>Répond difficilement aux sollicitations (obligatoire pour le score 3)</p> <p>N'initie pas l'interaction ou seulement en cas de détresse (obligatoire pour le score 3)</p> <p>Si semble initier (ex apporte un jouet), est en fait indifférent à la réponse de son interlocuteur</p> <p>Pas d'imitation</p> <p>Absence de mimique ou mimique figée ; rires / sourires non adressés</p> <p>Explore l'environnement, déambule, mais ne s'intéresse pas aux personnes</p>
	4	<p>Aucune réponse aux sollicitations d'interaction (obligatoire pour score 4)</p> <p>Ne regarde pas les visages ou regard « passe muraille »</p> <p>Activités solitaires et stéréotypées, déconnectées de l'environnement</p> <p>Absence de communication non verbale</p> <p>Pseudo surdité (ne réagit pas aux bruits forts)</p>

Tableau S3 (suite et fin)

Type de score	Valeur	Libellé
Score de sommeil	1	Pas de problème en général (ou petites difficultés très occasionnelles) S'endort facilement et dort très bien en général
	2	Difficultés d'endormissement ou de réveil occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents
Score d'alimentation	1	Pas de problème en général (ou petites difficultés très occasionnelles) Mange très bien en général
	2	Difficultés aux repas occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème, difficulté lors du passage à la diversification
	4	Difficultés importantes : réflexe nauséeux fréquent, préoccupation quotidienne à la limite du tolérable pour les parents – nutrition par sonde naso-gastrique ou gastrostomie
Score visuel	1	Bonne vision de chaque œil et vision des reliefs normale Fixation stable et poursuite oculaire normale sur chaque œil Si trouble de réfraction : modéré, ne nécessitant pas de correction Pas de gêne à l'occlusion
	2	Exophorie intermittente Trouble de réfraction nécessitant une correction Si amblyopie : corrigée par les lunettes Port de cache
	3	Trouble de réfraction sévère d'au moins 1 œil Amblyopie partiellement corrigée par les lunettes Strabisme manifeste Pas de vision binoculaire
	4	Amblyopie bilatérale Fixation absente ou très instable Poursuite oculaire absente ou très limitée
Score auditif	1	Aucune anomalie ; si explorations : normales
	2	Surdité unilatérale ou bilatérale légère (perte de 21 à 40dB)
	3	Surdité bilatérale moyenne (perte de 41 à 70 dB)
	4	Surdité bilatérale sévère ou profonde (perte > 70 dB)
Appréciation globale des parents	1	Tout va pour le mieux (si problèmes : bien acceptés et bien gérés)
	2	Contraintes modérées / inquiétudes occasionnelles ayant globalement peu d'impact sur l'équilibre familial
	3	Inquiétude et contraintes importantes. Impact notable sur la famille
	4	Impact très lourd, ayant totalement bouleversé la famille

Tableau S4 : Description des scores à 2 ans

Type de score	Valeur	Libellé
Score respiratoire	1	Aucun problème
	2	Bronchites / bronchiolites occasionnelles peu sévères (maxi 2)
	3	Bronchites / bronchiolites répétées (= 3) et/ou sévères ou wheezing fréquent
	4	Oxygénothérapie ou trachéo / hospitalisations multiples pour problèmes respiratoires
Autres problèmes de santé	1	Aucun problème
	2	Problèmes occasionnels facilement résolus
	3	Problèmes répétitifs ou non encore résolus
	4	Problèmes gênant la vie quotidienne ou nécessitant plusieurs hospitalisations
Score moteur	1	Marche fluide (obligatoire pour score 1) Absence de Déficit moteur d'un membre (obligatoire pour score 1) Court Monte et descend l'escalier debout Saute sur les 2 pieds Shoote dans un ballon
	2	Marche restant hésitante Chute non protégée Monte quelques marche mais ne descend pas Déficit moteur d'un membre
	3	Diplégie ou hémiplegie spastique Spasticité ou dystonie Démarche anormale, avec équin dynamique ; ou marche assistée Tient assis
	4	Quadriplégie spastique Pas de marche même assistée Ne Tient pas assis Préhension pathologique
Score de motricité fine / coordination	1	Préhension et manipulation fine aisée Utilise correctement la cuillère pour manger Emboîte et encastre avec réflexion et application Empile 3 cubes Bonne coordination bi manuelle (ex : tient le petit flacon d'une main pour introduire la pastille et visser le bouchon). Gribouille en portant intérêt aux traces laissées, voire commence à imiter un trait orienté Commence à se déshabiller et s'habiller (veut faire seul) Renverse pour vider
	2	Préhension fine normale mais manipulation plus grossière Utilise la cuillère mais renverse souvent Encastremets aléatoires Gribouille sans porter d'intérêt aux traces laissées, pas d'imitation Dépose un cube sur la tour mais ne construit pas de lui-même

Tableau S4 (suite)

Type de score	Valeur	Libellé
Score de motricité fine / coordination	3	Préhension grossière mais avec intentionnalité Saisit sa cuillère (mais ne parvient pas à manger seul) Non intéressé par les encastremets Intéressé par la tour mais ne dépose pas un cube Remet dans la boîte Intéressé par les jeux de cause à effet (ex : presse un bouton pour voir l'effet)
	4	Préhension grossière mais sans intentionnalité Absence d'intérêt pour les jeux de cause à effet Activités très stéréotypées : vide, jette, tape
Score de langage	1	Dit plus de 20 mots Associe 2 mots (hors mots-phrases type « il est là ») Très bonne compréhension verbale (comprend > 50 mots) Dénomme au moins 3 objets / images parmi un choix Désigne beaucoup d'images
	2	Dit environ 10 mots Jargon modulé Bonne compréhension verbale : peut désigner 1 objet / 1 image parmi un choix Très bonne compréhension contextuelle (va chercher, apporte, mets, enlève, ouvre, ferme...) en consultation
	3	Dit moins de 5 mots Compréhension contextuelle limitée aux actions du quotidien Compréhension contextuelle en consultation limitée aux actions simples (donne assieds-toi, debout...) Regarde les objets courants cités (ballon, chaussures...) mais ne désigne pas (sauf parties du visage/ du corps)
	4	Pas de mot (sauf mm ou équivalent pour maman, papa) Compréhension contextuelle observée seulement par les parents sur quelques actions très ritualisées du quotidien Pas de compréhension contextuelle observée en consultation Ne désigne pas les parties du corps / du visage
Score relationnel	1	Répond facilement aux sollicitations d'interaction et d'attention conjointe Maintient l'interaction Initie lui-même l'interaction Répond à son prénom Communication non verbale riche Mimique riche, adressée Imitations facilement observées

Tableau S4 (suite)

Type de score	Valeur	Libellé
Score relationnel	2	Répond aux sollicitations mais ne maintient pas l'interaction, poursuit le jeu en solo Initie peu l'interaction Comportement plus « indépendant » - ne cherche pas trop le contact physique Communication non verbale pauvre ; mimique pauvre mais adressée Peu d'imitations Bébé trop calme et trop sérieux ou au contraire agité
	3	Répond difficilement aux sollicitations N'initie pas l'interaction ou seulement en cas de détresse Si semble initier (ex apporte un jouet), est en fait indifférent à la réponse de son interlocuteur Pas d'imitation Absence de mimique ou mimique figée ; rires / sourires non adressés Explore l'environnement, déambule, mais ne s'intéresse pas aux personnes
	4	Aucune réponse aux sollicitations d'interaction Ne regarde pas les visages ou regard « passe muraille » Activités solitaires et stéréotypées, déconnectées de l'environnement Absence de communication non verbale Pseudo surdité (ne réagit pas aux bruits forts)
Score de sommeil	1	Pas de problème en général (ou petites difficultés très occasionnelles) S'endort facilement et dort très bien en général
	2	Difficultés d'endormissement ou de réveil occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents
Score de repas / alimentation	1	Pas de problème en général Mange très bien en général Peu ou pas de sélectivité
	2	Difficultés aux repas occasionnelles et/ou facilement résolues Refus de certains aliments
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème, difficulté lors du passage à la diversification Refuse encore les morceaux
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents – gavage par sonde naso-gastrique ou gastrostomie

Tableau S4 (suite et fin)

Type de score	Valeur	Libellé
Score de comportement psycho-affectif	1	Pas de problème en général (ou petites difficultés très occasionnelles) Enfant de bonne humeur et très « facile » en général
	2	Difficultés occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents
Score visuel	1	Bonne vision de chaque œil et vision des reliefs normale Fixation stable et poursuite oculaire normale sur chaque œil Si trouble de réfraction : modéré, ne nécessitant pas de correction Pas de gêne à l'occlusion
	2	Exophorie intermittente Trouble de réfraction nécessitant une correction Si amblyopie : corrigée par les lunettes Port de cache
	3	Trouble de réfraction sévère d'au moins 1 œil Amblyopie partiellement corrigée par les lunettes Strabisme manifeste Pas de vision binoculaire
	4	Amblyopie bilatérale Fixation absente ou très instable Poursuite oculaire absente ou très limitée
Score auditif	1	Aucune anomalie ; si explorations : normales
	2	Surdit� unilat�rale ou bilat�rale l�g�re (perte de 21 � 40dB)
	3	Surdit� bilat�rale moyenne (perte de 41 � 70 dB)
	4	Surdit� bilat�rale s�v�re ou profonde (perte > 70 dB)
Appr�ciation globale des parents	1	Tout va pour le mieux (si probl�mes : bien accept�s et bien g�r�s)
	2	Contraintes mod�r�es / inqui�tudes occasionnelles ayant globalement peu d'impact sur l'�quilibre familial
	3	Inqui�tude et contraintes importantes.
	4	Impact notable sur la famille
	4	Impact tr�s lourd, ayant totalement boulevers� la famille

Tableau S5 : Description des scores à 3 ans

Type de score	Valeur	Libellé
Score respiratoire	1	Aucun problème
	2	Crises d'asthme occasionnelles peu sévères (maxi 2)
	3	Crises d'asthme sévères ou répétées (= 3) et/ou sévères ou wheezing fréquent
	4	Oxygénothérapie ou trachéotomie ou pathologie respiratoires nécessitant des hospitalisations multiples
Autres problèmes de santé	1	Aucun problème
	2	Problèmes occasionnels facilement résolus
	3	Problèmes répétitifs ou non encore résolus
	4	Problèmes gênant la vie quotidienne ou nécessitant plusieurs hospitalisations
Score moteur	1	Pas de déficit d'un membre (obligatoire si score 1) Monte les escaliers en alternant, descend les escaliers de façon autonome. Tricycle, vélo avec stabilisateurs ou trottinette Marche et court avec aisance Shoote dans un ballon avec force et bon équilibre Peut attraper un gros ballon Sauter à pieds joints
	2	Marche et court avec moins d'aisance Monte les escaliers sans alterner ou descend seulement tenu Déficit moteur d'un membre
	3	Diplégie ou hémiplegie spastique Spasticité ou dystonie Démarche anormale, avec équin dynamique ; ou marche assistée Tient assis
	4	Quadriplégie spastique Pas de marche même assistée Ne Tient pas assis Préhension pathologique
Score de motricité fine / coordination	1	S'habille presque seul (sauf boutons) Lave et sèche ses mains Encastre toutes les formes avec anticipation et orientation correcte de l'objet Tour de plus de 8 cubes Puzzle de 6 pièces Dessine un rond fermé S'intéresse et participe au dessin symbolique (ex : bonhomme) voire le réalise seul
	2	Enlève ses vêtements, en enfile certains Encastre toutes les formes simples mais sans anticipation (procède par essais/erreurs) ou avec des difficultés Tour de 4 à 8 cubes Fait un rond non fermé, imite les traits orientés Participe au bonhomme par imitation mais sans symbolisme

Tableau S5 (suite)

Type de score	Valeur	Libellé
Score de motricité fine / coordination	3	Encastre seulement le rond Empile 2-3 cubes Gribouillis sans imitation Peu intéressé par les jeux de construction, préfère les jeux de cause à effet
	4	Niveau de jeu : purement sensoriel Aucun encastrement N'empile pas Ne tient pas le crayon ou ne s'en sert pas pour gribouiller Jette, tape, vide.
Score de langage	1	Phrases (sujet, verbe, complément) Utilise les pronoms je, tu, il, elle Décrit une action sur l'image : répond aux questions « Que fait... ? », « A quoi sert... ? » Compréhension syntaxique des petits mots avec, sur, sous, ...
	2	Associe les mots, phrases sujet verbe ou verbe objet Nomme les objets mais ne décrit pas une action Compréhension syntaxique faible : sur imagier ne répond qu'à la question « Qu'est-ce que c'est... ? »
	3	Mots et mots/phrases isolés Désigne quelques images ou objets Agrammatique ou inintelligible pour les non-familiers
	4	Moins de 10 mots Pas de désignation
	5	Non évaluable
Score relationnel	1	Répond facilement aux sollicitations d'interaction et d'attention conjointe, maintient l'interaction et initie lui-même l'interaction Communication verbale ou non verbale riche Manifestation d'empathie (va consoler un camarade) Jeux symboliques riches et jeux de rôle
	2	Répond aux sollicitations mais ne maintient pas l'interaction, poursuit le jeu en solo Initie peu l'interaction Comportement plus « indépendant » - ne cherche pas trop le contact physique Communication non verbale pauvre Intérêts sélectifs mais non envahissants Enfant trop calme et trop sérieux ou au contraire agité Jeux symboliques pauvres Participe au bonhomme par imitation mais sans symbolisme

Tableau S5 (suite)

Type de score	Valeur	Libellé
Score relationnel	3	Répond difficilement aux sollicitations N'initie pas l'interaction ou seulement en cas de détresse Si semble initier (ex apporte un jouet), est en fait indifférent à la réponse de son interlocuteur Pas d'imitation Absence de mimique ou mimique figée ; rires / sourires non adressés Explore l'environnement, déambule, mais ne s'intéresse pas aux personnes
	4	Aucune réponse aux sollicitations d'interaction Ne regarde pas les visages ou regard « passe muraille » Activités solitaires et stéréotypées, déconnectées de l'environnement Absence de communication non verbale Pseudo surdité (ne réagit pas aux bruits forts)
Score d'attention	1	Attention soutenue, perspicacité
	2	Attention plus labile, passe vite d'une activité à l'autre
	3	Gros troubles de concentration pour les activités de manipulation et les activités exploratoires
	4	Gros troubles de concentration pour toutes les activités
Score de sommeil	1	Pas de problème en général (ou petites difficultés très occasionnelles) S'endort facilement et dort très bien en général
	2	Difficultés d'endormissement ou de réveil occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents
Score d'alimentation	1	Pas de problème en général (ou petites difficultés très occasionnelles) Mange très bien en général
	2	Difficultés aux repas occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème, difficulté lors du passage à la diversification Repas très long Refus de la nouveauté Stratégie pour éviter déglutition et/ou mastication
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents – gavage par sonde naso-gastrique ou gastrostomie Problème de socialisation (ne peut manger qu'à la maison)

Tableau S5 (suite et fin)

Type de score	Valeur	Libellé
Score de comportement / sociabilité	1	Pas de problème en général (ou petites difficultés très occasionnelles) Enfant de bonne humeur et très « facile » en général
	2	Difficultés occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents
Score visuel	1	Bonne vision de chaque œil et vision des reliefs normale Fixation stable et poursuite oculaire normale sur chaque œil Si trouble de réfraction : modéré, ne nécessitant pas de correction Pas de gêne à l'occlusion
	2	Exophorie intermittente Trouble de réfraction nécessitant une correction Si amblyopie : corrigée par les lunettes Port de cache
	3	Trouble de réfraction sévère d'au moins 1 œil Amblyopie partiellement corrigée par les lunettes Strabisme manifeste Pas de vision binoculaire
	4	Amblyopie bilatérale Fixation absente ou très instable Poursuite oculaire absente ou très limitée
Score auditif	1	Aucune anomalie ; si explorations : normales
	2	Surdit� unilat�rale ou bilat�rale l�g�re (perte de 21 � 40dB)
	3	Surdit� bilat�rale moyenne (perte de 41 � 70 dB)
	4	Surdit� bilat�rale s�v�re ou profonde (perte > 70 dB)
Appr�ciation globale des parents	1	Tout va pour le mieux (si probl�mes : bien accept�s et bien g�r�s)
	2	Contraintes mod�r�es / inqui�tudes occasionnelles ayant globalement peu d'impact sur l'�quilibre familial
	3	Inqui�tude et contraintes importantes Impact notable sur la famille
	4	Impact tr�s lourd, ayant totalement boulevers� la famille

Tableau S6 : Description des scores à 4 ans

Type de score	Valeur	Libellé
Score respiratoire	1	Aucun problème
	2	Crises d'asthme occasionnelles peu sévères
	3	Crises d'asthme sévères ou répétées (= 3) ou wheezing fréquent
	4	Oxygénothérapie ou trachéotomie ou pathologie respiratoire nécessitant des hospitalisations multiples
Autres problèmes de santé	1	Aucun problème
	2	Problèmes occasionnels facilement résolus ou problème ne gênant pas la vie quotidienne et ne nécessitant pas de consultations fréquentes
	3	Problèmes répétitifs ou non encore résolus ou nécessitant des consultations répétées
	4	Problèmes gênant la vie quotidienne ou nécessitant plusieurs hospitalisations
Score moteur	1	Pas de déficit d'un membre (obligatoire si score 1) Monte et descend les escaliers en alternant Sait pédaler Saute d'une marche à pieds joints sans appréhension Cloche pied sur le pied dominant Lance le ballon de façon dirigée ; attrape le ballon avec succès le plus souvent A l'aise dans les jeux moteurs de l'école/centre de loisir
	2	Court avec moins d'aisance, trébuche souvent – fatigabilité à la marche Manque d'aisance pour sauter Lance le ballon avec peu de force et de façon non ciblée ; réception difficile, peu coordonnée Déficit d'un membre
	3	Diplégie ou hémiplegie spastique Spasticité ou dystonie Démarche anormale, avec équin dynamique ; ou marche assistée Tient assis
	4	Quadriplégie spastique Pas de marche même assistée Ne Tient pas assis Préhension pathologique

Tableau S6 (suite)

Type de score	Valeur	Libellé
Fonctions non verbales	1	<p>Autonomie : s'habille pratiquement seul, met les gros boutons et utilise la fermeture éclair, va seul aux toilettes, mange seul proprement avec cuillère et fourchette</p> <p>Constructions : reproduit aisément le pont avec les cubes, fait aisément et avec adresse une tour de 10 cubes, reproduit la « locomotive » orientée, la pyramide de 6 cubes</p> <p>Dénombrer 3-4 cubes</p> <p>Graphisme : dessine un bonhomme (si encore têtard ajoute des détails (doigts, pieds, mains...), fait un rond bien rond et fermé, les croix orientées (+ x) et le carré, copie le dessin des lunettes et peut le refaire de mémoire ; figures fermées de spatialisation. Repère les parallèles et les obliques.</p> <p>Encastrement : toutes les formes très aisément avec anticipation</p>
	2	<p>Autonomie : enlève ses vêtements, en enfile certains, moins d'autonomie</p> <p>Constructions : moins d'adresse</p> <p>Ne dénombre pas</p> <p>Graphisme : bonhomme têtard rudimentaire, dessine le rond, la croix plus difficilement (et non orientées), pas le carré</p> <p>Encastrement : toutes les formes mais avec des défauts d'anticipation et d'orientation, procède le plus souvent par essais/erreurs</p>
	3	<p>Encastre seulement le rond aisément</p> <p>Tour de quelques cubes</p> <p>Graphisme : pas de bonhomme, trace des traits (imite) et des ronds non fermés ou irréguliers</p>
	4	<p>Activités très pauvres et stéréotypées</p> <p>Encastrement très difficiles ou absents, n'empile pas</p> <p>Graphisme : gribouillis sans imitation</p>
Langage	1	<p>Fait des énoncés de 4 à 6 mots avec verbe ; raconte ce qu'il fait</p> <p>Conjugué les verbes ; utilise tous les pronoms</p> <p>Comprend et utilise : sur, sous, devant, derrière</p> <p>Connaît le nom des objets domestiques courants mais habituellement non utilisés par l'enfant (radiateur, poêle...)</p> <p>Utilise différents temps (présent, passé, futur proche)</p> <p>Début du sens de la temporalité (matin, soir)</p> <p>Répond à : que fais-tu quand tu as froid, faim, quand tu es fatigué ?</p> <p>Répond aux questions « Pourquoi ? »</p> <p>Participe facilement aux conversations, pose des questions de type « Pourquoi ? »</p> <p>Parfaitement intelligible, si défaut articulatoire : mineur (ex : zozote)</p>

Tableau S6 (suite)

Type de score	Valeur	Libellé
Langage	2	Phrases simples (sujet, verbe, complément) Exprime une action de l'image Vocabulaire spatial peu maîtrisé ; vocabulaire peu riche ; Erreurs phonologiques encore fréquentes (améliorables par répétition) Compréhension syntaxique partielle (ne saisit pas une question avec « Pourquoi ?, Avec quoi ? ... »)
	3	Associe les mots sans syntaxe, répète Nomme quelques objets, désigne quelques images Ne comprend que les questions élémentaires ; sinon répond par oui ou non ou le dernier mot de la question Description d'images peu informative Ou difficultés articulatoires (peu améliorées par la répétition)
	4	Pas de mots ou quelques mots isolés ou inintelligibles
	5	Non évaluable (mutisme irréductible)
	Score relationnel	1
	2	Répond aux sollicitations mais ne maintient pas l'interaction (poursuit le jeu en solo) Initie peu l'interaction, préfère jouer seul Ne cherche pas trop le contact physique ni le contact avec ses pairs Intérêts sélectifs mais non envahissants Enfant trop calme et trop sérieux ou au contraire agité Ne participe pas aux jeux de rôle
	3	Répond difficilement aux sollicitations N'initie pas l'interaction ou seulement en cas de détresse Si semble initier (ex apporte un jouet), est en fait indifférent à la réponse de son interlocuteur Communication non verbale pauvre Absence de mimique ou mimique figée ; rires / sourires non adressés Explore l'environnement, déambule, mais ne s'intéresse pas aux personnes Intérêts sélectifs très restreints, non partagés, envahissants
	4	Aucune réponse aux sollicitations d'interaction Ne regarde pas les visages ou regard « passe muraille » Activités solitaires et stéréotypées, déconnectées de l'environnement Absence de communication non verbale Pseudo surdité (ne réagit pas aux bruits forts)

Tableau S6 (suite)

Type de score	Valeur	Libellé
Score d'attention	1	Attention soutenue, perspicacité
	2	Attention plus labile, passe vite d'une activité à l'autre
	3	Gros troubles de concentration pour les activités de manipulation et les activités exploratoires
	4	Gros troubles de concentration pour toutes les activités
Score de sommeil	1	Pas de problème en général (ou petites difficultés très occasionnelles) S'endort facilement et dort très bien en général
	2	Difficultés d'endormissement ou de réveil occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents
Score d'alimentation	1	Pas de problème en général (ou petites difficultés très occasionnelles) Mange très bien en général
	2	Difficultés aux repas occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème, difficulté lors du passage à la diversification Repas très long Refus de la nouveauté Stratégie pour éviter déglutition et/ou mastication
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents – gavage par sonde nasogastrique ou gastrostomie Problème de socialisation (ne peut manger qu'à la maison)
Score de comportement psycho-affectif	1	Pas de problème en général (ou petites difficultés très occasionnelles) Enfant de bonne humeur et très « facile » en général
	2	Difficultés occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents
Score visuel	1	Bonne vision de chaque œil et vision des reliefs normale Fixation stable et poursuite oculaire normale sur chaque œil Si trouble de réfraction : modéré, ne nécessitant pas de correction Pas de gêne à l'occlusion
	2	Exophorie intermittente Trouble de réfraction nécessitant une correction Si amblyopie : corrigée par les lunettes Port de cache

Tableau S6 (suite et fin)

Type de score	Valeur	Libellé
Score visuel	3	Trouble de réfraction sévère d'au moins 1 œil Amblyopie partiellement corrigée par les lunettes Strabisme manifeste Pas de vision binoculaire
	4	Amblyopie bilatérale Fixation absente ou très instable Poursuite oculaire absente ou très limitée
Score auditif	1	Aucune anomalie ; si explorations : normales
	2	Surdit� unilat�rale ou bilat�rale l�g�re (perte de 21 � 40dB)
	3	Surdit� bilat�rale moyenne (perte de 41 � 70 dB)
	4	Surdit� bilat�rale s�v�re ou profonde (perte > 70 dB)
Appr�ciation globale des parents	1	Tout va pour le mieux (si probl�mes : bien accept�s et bien g�r�s)
	2	Contraintes mod�r�es / inqui�tudes occasionnelles ayant globalement peu d'impact sur l'�quilibre familial
	3	Inqui�tude et contraintes importantes. Impact notable sur la famille
	4	Impact tr�s lourd, ayant totalement boulevers� la famille

Tableau S7 : Description des scores à 5 ans

Type de score	Valeur	Libellé
Score respiratoire	1	Aucun problème
	2	Bronchites / toux / wheezing occasionnels peu sévères
	3	Bronchites / toux / wheezing sévères ou répétées (= 3 / an) Ou asthme justifiant un traitement de fond
	4	Asthme sévère, insuffisance respiratoire ou autre pathologie nécessitant des hospitalisations multiples Oxygénothérapie ou trachéotomie
Autres problèmes de santé	1	Aucun problème
	2	Problèmes occasionnels facilement résolus ou problèmes ne gênant pas la vie quotidienne et ne nécessitant pas de consultations fréquentes
	3	Problèmes répétitifs ou non encore résolus ou nécessitant des consultations répétées ou des consultations spécialisées rapprochées (= 3/ an)
	4	Problèmes gênant la vie quotidienne ou nécessitant plusieurs hospitalisations
Score moteur	1	Pas de déficit d'un membre (obligatoire si score 1) Monte et descend les escaliers en alternant Sait pédaler Saute d'une marche à pieds joints sans appréhension Cloche pied sur le pied dominant Lance le ballon de façon dirigée ; attrape le ballon avec succès le plus souvent A l'aise dans les jeux moteurs de l'école/centre de loisirs
	2	Court avec moins d'aisance, trébuche souvent – fatigabilité à la marche Manque d'aisance pour sauter Lance le ballon avec peu de force et de façon non ciblée ; réception difficile, peu coordonnée Déficit d'un membre
	3	Diplégie ou hémiplégie spastique Spasticité ou dystonie Démarche anormale, avec équin dynamique ; ou marche assistée Tient assis
	4	Quadriplégie spastique Pas de marche même assistée Ne Tient pas assis Préhension pathologique
Fonctions non verbales	1	Autonomie : s'habille complètement seul ; autonome aux toilettes ; mange seul (sait se servir du couteau) Constructions : reproduit l'escalier de 6 cubes ; constructions en 3D sur modèle Dénombrer plus de 5 cubes sans erreur de pointage, notion de cardinalité acquise Graphisme : dessine un bonhomme complet avec détails ; reproduit un rectangle avec diagonales, un triangle, assemblage de figures géométriques simples ; écrit son prénom en lettres bâton

Tableau S7 (suite)

Type de score	Valeur	Libellé
Fonctions non verbales	2	Autonomie : s'habille en partie seul (aidé pour les boutons, les collants...); moins d'autonomie; inverse endroit/envers, devant-derrière Constructions : escalier et constructions 3D avec modèle et aide verbale - procède le plus souvent par essais/erreurs Dénombrement : erreurs de comptine ou de pointage Graphisme : bonhomme immature; dessine le carré; doit être aidé par des repères pour les obliques / les diagonales; figures fermées de spatialisation réussies
	3	Echec malgré plusieurs essais, aide verbale ou facilitation topographique des 3 éléments (construction, graphisme, dénombrement,) cités en score 2 Si difficultés graphiques, construction du carré pas mieux réussie par assemblage de 4 crayons
	4	Activités très pauvres et stéréotypées
	5	Non évaluable : opposition irréductible ou difficultés importantes dans un autre domaine (motricité, vision...)
	Score de langage	1
	2	Langage avec moins de spontanéité, moins informatif Vocabulaire moins riche Syntaxe moins élaborée Erreurs phonologiques persistantes
	3	Associe les mots; phrases élémentaires Ou difficultés articulatoires
	4	Au mieux nomme quelques objets, désigne quelques images Pas de mots ou quelques mots isolés ou inintelligibles
	5	Non évaluable (mutisme irréductible)
Score relationnel	1	Répond facilement aux sollicitations d'interaction et d'attention conjointe, maintient l'interaction et initie lui-même l'interaction (obligatoire pour score 1) Communication verbale ou non verbale riche Manifestation d'empathie (va consoler un camarade) Jeux de rôle avec expression d'affects Jeux interactifs riches Participe aux jeux d'humour des enfants de son âge

Tableau S7 (suite)

Type de score	Valeur	Libellé
Score relationnel	2	Répond aux sollicitations mais ne maintient pas l'interaction, poursuit le jeu en solo (obligatoire pour score 2) Initie peu l'interaction, préfère jouer seul Ne cherche pas trop le contact physique ni le contact avec ses pairs Intérêts sélectifs mais non envahissants Enfant trop calme et trop sérieux ou au contraire agité Ne participe pas aux jeux de rôle
	3	Répond difficilement aux sollicitations (obligatoire pour score 3) N'initie pas l'interaction ou seulement en cas de détresse Si semble initier (ex apporte un jouet), est en fait indifférent à la réponse de son interlocuteur Communication non verbale pauvre Absence de mimique ou mimique figée ; rires / sourires non adressés Explore l'environnement, déambule, mais ne s'intéresse pas aux personnes Intérêts sélectifs très restreints, non partagés, envahissants
	4	Aucune réponse aux sollicitations d'interaction (obligatoire pour score 4) Ne regarde pas les visages ou regard « passe muraille » Activités solitaires et stéréotypées, déconnectées de l'environnement Absence de communication non verbale Pseudo surdité (ne réagit pas aux bruits forts)
Score d'attention	1	Attention soutenue Suit bien les consignes
	2	Attention plus labile Se laisse facilement distraire
	3	Attention très labile ; abandonne vite une activité ; très brouillon Interrompt souvent les autres - impulsivité Remue beaucoup
	4	Déficit attentionnel marqué avec hyperactivité motrice et impulsivité altérant la vie quotidienne et les apprentissages scolaires
	5	Non évaluable (ex. déficit cognitif sévère)
Score de sommeil	1	Pas de problème en général (ou petites difficultés très occasionnelles) S'endort facilement et dort très bien en général
	2	Difficultés d'endormissement ou de réveil occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents

Tableau S7 (suite)

Type de score	Valeur	Libellé
Score d'alimentation	1	Pas de problème en général (ou petites difficultés très occasionnelles) Mange très bien en général
	2	Difficultés aux repas occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème, difficulté lors du passage à la diversification Repas très long Refus de la nouveauté Stratégie pour éviter déglutition et/ou mastication
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents – gavage par sonde naso-gastrique ou gastrostomie Problème de socialisation (ne peut manger qu'à la maison)
Score de comportement psycho-affectif	1	Pas de problème en général (ou petites difficultés très occasionnelles) Enfant de bonne humeur et très « facile » en général
	2	Difficultés occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents
Score visuel	1	Bonne vision de chaque œil et vision des reliefs normale Fixation stable et poursuite oculaire normale sur chaque œil Si trouble de réfraction : modéré, ne nécessitant pas de correction Pas de gêne à l'occlusion
	2	Exophorie intermittente Trouble de réfraction nécessitant une correction Si amblyopie : corrigée par les lunettes Port de cache
	3	Trouble de réfraction sévère d'au moins 1 œil Amblyopie partiellement corrigée par les lunettes Strabisme manifeste Pas de vision binoculaire
	4	Amblyopie bilatérale Fixation absente ou très instable Poursuite oculaire absente ou très limitée
Score auditif	1	Aucune anomalie ; si explorations : normales
	2	Surdit� unilat�rale ou bilat�rale l�g�re (perte de 21 � 40dB)
	3	Surdit� bilat�rale moyenne (perte de 41 � 70 dB)
	4	Surdit� bilat�rale s�v�re ou profonde (perte > 70 dB)

Tableau S7 (suite et fin)

Type de score	Valeur	Libellé
Appréciation globale des parents	1	Tout va pour le mieux (si problèmes : bien acceptés et bien gérés)
	2	Contraintes modérées / inquiétudes occasionnelles ayant globalement peu d'impact sur l'équilibre familial
	3	Inquiétude et contraintes importantes Impact notable sur la famille
	4	Impact très lourd, ayant totalement bouleversé la famille

Tableau S8 : Description des scores à 6 ans

Type de score	Valeur	Libellé
Score respiratoire	1	Aucun problème
	2	Bronchites / toux / wheezing occasionnels peu sévères
	3	Bronchites / toux / wheezing sévères ou répétées (= 3 / an) ou asthme justifiant un traitement de fond
	4	Asthme sévère, insuffisance respiratoire ou autre pathologie nécessitant des hospitalisations multiples Oxygénothérapie ou trachéotomie
Autres problèmes de santé	1	Aucun problème
	2	Problèmes occasionnels facilement résolus ou problème ne gênant pas la vie quotidienne et ne nécessitant pas de consultations fréquentes
	3	Problèmes répétitifs ou non encore résolus ou nécessitant des consultations répétées ou des consultations spécialisées rapprochées (= 3/ an)
	4	Problèmes gênant la vie quotidienne ou nécessitant plusieurs hospitalisations
Score moteur	1	Pas de déficit d'un membre Monte et descend les escaliers en alternant Sait pédaler Saute d'une marche à pieds joints sans appréhension Cloche pied sur le chaque pied Tient en équilibre sur le pied dominant Lance le ballon de façon dirigée ; attrape le ballon avec succès le plus souvent A l'aise dans les jeux moteurs de l'école/centre de loisir
	2	Court avec moins d'aisance, trébuche souvent – fatigabilité à la marche Manque d'aisance pour sauter Peu d'aisance dans les jeux sportifs Lance le ballon avec peu de force et de façon non ciblée ; réception difficile, peu coordonnée Déficit d'un membre
	3	Diplégie ou hémiplegie spastique Spasticité ou dystonie Démarche anormale, avec équin dynamique ; ou marche assistée Tient assis
	4	Quadriplégie spastique Pas de marche même assistée Ne Tient pas assis Préhension pathologique

Tableau S8 (suite)

Type de score	Valeur	Libellé
Fonctions non verbales	1	Autonomie : s'habille complètement seul (sauf lacets) ; autonome aux toilettes ; mange seul (sait se servir du couteau, sait tartiner) ; se coiffe ; se mouche ; connaît sa droite et sa gauche Constructions : constructions en 3D sur modèle Dénombrer plus de 10 cubes sans erreur de pointage, notion de cardinalité acquise ; surcomptage de 1 et de 2 ; comptine des chiffres à rebours et décomptage de 1 Mémoire de travail : peut répéter 2 chiffres à l'envers Graphisme : dessin figuratif riche, avec détails ; reproduit les figures géométriques avec obliques et croisements (triangle, astérisque, diagonales) et un assemblage de figures géométriques (pseudo figure de Rey) ; début des lettres cursives ou reproduit des boucles et des boucles alternées (graphotomes) ; frise de Luria respectée
	2	Autonomie : s'habille en partie seul (aidé pour les boutons, les collants...); moins d'autonomie ; inverse endroit/envers, devant-derrrière Constructions : escalier et constructions 3D avec modèle et aide verbale - procède le plus souvent par essais/erreurs Dénombrement : erreurs de comptine ou de pointage ; erreurs de comptine ou de pointage ; pas de surcomptage Mémoire de travail : ne restitue pas 2 chiffres à l'envers Graphisme : bonhomme immature ; dessine le carré ; doit être aidé par des repère pour les obliques / les diagonales ; figures fermées de spatialisation réussies
	3	Echec malgré plusieurs essais, aide verbale ou facilitation topographique des 3 éléments (construction, graphisme, dénombrement) cités en score 2 Si difficultés graphiques, construction du carré pas mieux réussie par assemblage de 4 crayons
	4	Activités très pauvres et stéréotypées
	5	Non évaluable : opposition irréductible ou difficultés importantes dans un autre domaine (motricité, vision...)
Langage	1	Langage spontané riche, informatif – peut raconter un fait vécu Utilise ou comprend une syntaxe élaborée (ex. propositions relatives) ; respect des accords noms-adjectifs Utilise le langage des sentiments et émotions (peur / tristesse / colère...) Définit certains mots (définitions et non attributs) - Vocabulaire des contraires Début des fluences (peut citer plusieurs mots d'une catégorie) Tous les sons maîtrisés (sauf éventuellement ch/j et s/z)
	2	Langage avec moins de spontanéité, moins informatif Vocabulaire moins riche Syntaxe moins élaborée Erreurs phonologiques persistantes

Tableau S8 (suite)

Type de score	Valeur	Libellé
Langage	3	Associe les mots ; phrases élémentaires Ou difficultés articulatoires
	4	Au mieux nomme quelques objets, désigne quelques images Pas de mots ou quelques mots isolés ou inintelligibles
	5	Non évaluable (mutisme irréductible)
Pré-requis du langage écrit	1	Segmentation des syllabes Soustraction de syllabes Reconnaissance des rimes
	2	Seulement segmentation de syllabes (pas de soustraction)
	3	Pas de segmentation des syllabes
	4	Non explorable
Score relationnel	1	Répond facilement aux sollicitations d'interaction et d'attention conjointe, maintient l'interaction et initie lui-même l'interaction (obligatoire pour score 1) Communication verbale ou non verbale riche Manifestation d'empathie (va consoler un camarade) Jeux de rôle avec expression d'affects Jeux interactifs riches Participe aux jeux d'humour des enfants de son âge
	2	Répond aux sollicitations mais ne maintient pas l'interaction, poursuit le jeu en solo (obligatoire pour score 2) Initie peu l'interaction, préfère jouer seul Ne cherche pas trop le contact physique ni le contact avec ses pairs Intérêts sélectifs mais non envahissants Enfant trop calme et trop sérieux ou au contraire agité Ne participe pas aux jeux de rôle
	3	Répond difficilement aux sollicitations (obligatoire pour score 3) N'initie pas l'interaction ou seulement en cas de détresse Si semble initier (ex apporte un jouet), est en fait indifférent à la réponse de son interlocuteur Communication non verbale pauvre Absence de mimique ou mimique figée ; rires / sourires non adressés Explore l'environnement, déambule, mais ne s'intéresse pas aux personnes Intérêts sélectifs très restreints, non partagés, envahissants
	4	Aucune réponse aux sollicitations d'interaction (obligatoire pour score 4) Ne regarde pas les visages ou regard « passe muraille » Activités solitaires et stéréotypées, déconnectées de l'environnement Absence de communication non verbale Pseudo surdité (ne réagit pas aux bruits forts)

Tableau S8 (suite)

Type de score	Valeur	Libellé
Score d'attention	1	Attention soutenue Suit bien les consignes
	2	Attention plus labile Se laisse facilement distraire
	3	Attention très labile ; abandonne vite une activité ; très brouillon Interrompt souvent les autres - impulsivité Remue beaucoup
	4	Déficit attentionnel marqué avec hyperactivité motrice et impulsivité altérant la vie quotidienne et les apprentissages scolaires
	5	Non évaluable (ex. déficit cognitif sévère)
Score de sommeil	1	Pas de problème en général (ou petites difficultés très occasionnelles) S'endort facilement et dort très bien en général
	2	Difficultés d'endormissement ou de réveil occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents
Score de repas / alimentation	1	Pas de problème en général (ou petites difficultés très occasionnelles) Mange très bien en général
	2	Difficultés aux repas occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème, difficulté lors du passage à la diversification Repas très long Refus de la nouveauté Stratégie pour éviter déglutition et/ou mastication
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents – gavage par sonde naso-gastrique ou gastrostomie Problème de socialisation (ne peut manger qu'à la maison)
Score de comportement / sociabilité	1	Pas de problème en général (ou petites difficultés très occasionnelles) Enfant de bonne humeur et très « facile » en général
	2	Difficultés occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents

Tableau S8 (suite et fin)

Type de score	Valeur	Libellé
Score visuel	1	Bonne vision de chaque œil et vision des reliefs normale Fixation stable et poursuite oculaire normale sur chaque œil Si trouble de réfraction : modéré, ne nécessitant pas de correction Pas de gêne à l'occlusion
	2	Exophorie intermittente Trouble de réfraction nécessitant une correction Si amblyopie : corrigée par les lunettes Port de cache
	3	Trouble de réfraction sévère d'au moins 1 œil Amblyopie partiellement corrigée par les lunettes Strabisme manifeste Pas de vision binoculaire
	4	Amblyopie bilatérale Fixation absente ou très instable Poursuite oculaire absente ou très limitée
Score auditif	1	Aucune anomalie ; si explorations : normales
	2	Surdit� unilat�rale ou bilat�rale l�g�re (perte de 21 � 40 dB)
	3	Surdit� bilat�rale moyenne (perte de 41 � 70 dB)
	4	Surdit� bilat�rale s�v�re ou profonde (perte > 70 dB)
Score d'appr�ciation globale des parents	1	Tout va pour le mieux (si probl�mes : bien accept�s et bien g�r�s)
	2	Contraintes mod�r�es / inqui�tudes occasionnelles ayant globalement peu d'impact sur l'�quilibre familial
	3	Inqui�tude et contraintes importantes. Impact notable sur la famille
	4	Impact tr�s lourd, ayant totalement boulevers� la famille

Tableau S9 : Description des scores à 7 ans

Type de score	Valeur	Libellé
Score respiratoire	1	Aucun problème
	2	Bronchites / toux / wheezing occasionnels peu sévères
	3	Bronchites / toux / wheezing sévères ou répétées (= 3 / an) Ou asthme justifiant un traitement de fond
	4	Asthme sévère, insuffisance respiratoire ou autre pathologie nécessitant des hospitalisations multiples Oxygénothérapie ou trachéotomie
Autres problèmes de santé	1	Aucun problème
	2	Problèmes occasionnels facilement résolus ou problème ne gênant pas la vie quotidienne et ne nécessitant pas de consultations fréquentes
	3	Problèmes répétitifs ou non encore résolus ou nécessitant des consultations répétées ou des consultations spécialisées rapprochées (= 3/ an)
	4	Problèmes gênant la vie quotidienne ou nécessitant plusieurs hospitalisations
Score moteur	1	Pas de déficit d'un membre (obligatoire si score 1) Monte et descend les escaliers en alternant Sait pédaler Saute d'une marche à pieds joints sans appréhension Cloche pied sur le pied dominant Tiens en équilibre sur le pied dominant Lance le ballon de façon dirigée ; attrape le ballon avec succès le plus souvent A l'aise dans les jeux moteurs de l'école/centre de loisir
	2	Court avec moins d'aisance, trébuche souvent – fatigabilité à la marche Manque d'aisance pour sauter Vélo avec stabilisateur Lance le ballon avec peu de force et de façon non ciblée ; réception difficile, peu coordonnée Déficit d'un membre
	3	Diplégie ou hémiplegie spastique Spasticité ou dystonie Démarche anormale, avec équin dynamique ; ou marche assistée Tient assis
	4	Quadriplégie spastique Pas de marche même assistée Ne Tient pas assis Préhension pathologique

Tableau S9 (suite)

Type de score	Valeur	Libellé
Fonctions non verbales	1	<p>Autonomie : s'habille complètement seul (lacets, jeans ou autres vêtements étroits, petits boutons) ; complètement autonome à table (peut couper sa viande, verser adroitement de l'eau dans un verre)</p> <p>Constructions : reproduit des figures complexes en 3D</p> <p>Dénombrer : surcomptage et décomptage bien maîtrisés ; compte jusqu'à 60 ; comptine de 2 en 2 ; additionne des nombres de somme > 10 ; sait résoudre une soustraction simple ou une addition à trou ; sait faire des additions avec retenue (si CE1) ; permanence du nombre acquise. A le concept 1 dizaine = 10 unités. Lecture des nombres : lit les dizaines</p> <p>Graphisme : reproduit le losange, un assemblage complexe de figures géométriques ; écriture cursive harmonieuse</p>
	2	<p>Autonomie : s'habille en partie seul (aidé pour les boutons, les zips...) ; moins d'autonomie ou d'adresse à table</p> <p>Constructions : après plusieurs essais ou avec aide verbale pas à pas</p> <p>Dénombrement de 10 cubes correct, notion de cardinal acquis mais pas de permanence du nombre ; peut faire une addition en manipulant mais difficilement pour les sommes > 10 ; pas de soustraction</p> <p>Graphisme : reproduction approximative des figures complexes ; ou dysgraphie importante (peu lisible ou très coûteuse en énergie)</p>
	3	<p>Echec malgré plusieurs essais et aide verbale d'au moins 2 des 3 éléments (construction, graphisme, dénombrement,) cités en (2)</p> <p>Constructions : tour, éventuellement pont de 3 cubes</p> <p>Graphisme : dessins figuratifs élémentaires, peut tracer qq figures géométriques simples (ex carré, cercles) mais ne reproduit pas les associations de figures ; écrit juste son prénom</p> <p>Dénombrer : associe pointage et comptine mais ne dénombre pas > 5</p> <p>Beaucoup de persévérations</p>
	4	<p>Constructions absentes ou élémentaires (empile moins de 5 cubes)</p> <p>Graphisme absent ou élémentaire (pas de dessin figuratif ni figure fermée)</p> <p>Concepts numériques absents</p> <p>Ou activités très pauvres et stéréotypées</p>
	5	<p>Non évaluable : opposition irréductible ou difficultés importantes dans un autre domaine (motricité, vision...)</p>

Tableau S9 (suite)

Type de score	Valeur	Libellé
Langage oral	1	Langage spontané riche, informatif – peut raconter un fait vécu Utilise ou comprend une syntaxe élaborée (ex. propositions relatives) ; respect des accords noms-adjectifs Utilise le langage des sentiments et émotions (peur / tristesse / colère...) Définit certains mots (définitions et non attributs) - Vocabulaire des contraires Début des fluences (peut citer plusieurs mots d'une catégorie) Tous les sons maîtrisés (sauf éventuellement ch/j et s/z)
	2	Langage avec moins de spontanéité, moins informatif Vocabulaire moins riche Syntaxe moins élaborée Erreurs phonologiques persistantes
	3	Associe les mots ; phrases élémentaires Ou difficultés articulatoires
	4	Au mieux nomme qq objets, désigne qq images Pas de mots ou quelques mots isolés ou inintelligible
	5	Non évaluable (mutisme irréductible)
Langage écrit	1	Lit de façon assez fluide, retient et comprend ce qu'il lit Peut lire seul un livre de son niveau Ecrit correctement une phrase avec des mots simples sous la dictée
	2	Déchiffre mais lecture encore peu fluide (pas assez aisée pour avoir le plaisir de lire seul) Dictée : qq erreurs ou oublis de lettres Soustraction de syllabes possible
	3	Déchiffrage laborieux (bute sur chaque mot) ; beaucoup d'erreurs de décryptage Dictée : beaucoup d'erreurs ou omissions de lettres ; difficultés de segmentation des mots Segmentation en syllabes possible mais soustraction des syllabes absente ou difficile, identifie mal les rimes et les sons
	4	Peu ou pas d'acquisition en lecture Aucune capacité métaphonologique
	5	Non évaluable
Développement psychique et relationnel	1	Jeux de rôle avec expression d'affects Jeux interactifs riches Participe aux jeux d'humour des enfants de son âge Comprend les émotions de l'autre (ex. va consoler un camarade qui s'est blessé) Peut restituer un petit scénario issu d'un conte ou d'une histoire
	2	Eléments précités en (1) douteux ou inconstants Comportement plus « indépendant » - ne cherche pas trop le contact Peu d'intérêt pour les jeux interactifs et les jeux symboliques Enfant trop calme et trop sérieux ou au contraire agité Réactions excessives à certains bruits ou certains touchers

Tableau S9 (suite)

Type de score	Valeur	Libellé
Développement psychique et relationnel	3	Interactions fugaces ; regard volontiers fuyant Jeux dénués d'affects Communication verbale et non verbale limitée Confusion je / tu Eléments cités en (2) tous présents et marqués Pseudo surdité (ne réagit pas aux bruits forts) Eléments associés : phobies, angoisse, insomnie, anorexie
	4	Critères du DSM IV concernant les troubles envahissants du développement (troubles des interactions sociales ; altération qualitative de la communication ; caractère restreint, répétitif et stéréotypé des comportements, des intérêts et des activités)
Score d'attention	1	Attention soutenue Suit bien les consignes
	2	Attention plus labile Se laisse facilement distraire
	3	Attention très labile ; abandonne vite une activité ; très brouillon Interrompt souvent les autres - impulsivité Remue beaucoup
	4	Déficit attentionnel marqué avec hyperactivité motrice et impulsivité altérant la vie quotidienne et les apprentissages scolaires
	5	Non évaluable (ex. déficit cognitif sévère)
Score de sommeil	1	Pas de problème en général (ou petites difficultés très occasionnelles) S'endort facilement et dort très bien en général
	2	Difficultés d'endormissement ou de réveil occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents
Score d'alimentation	1	Pas de problème en général (ou petites difficultés très occasionnelles) Mange très bien en général
	2	Difficultés aux repas occasionnelles et/ou facilement résolues
	3	Difficultés fréquentes demandant une implication particulière des parents pour gérer le problème Repas très long Refus de la nouveauté Stratégie pour éviter déglutition et/ou mastication
	4	Difficultés importantes : préoccupation quotidienne à la limite du tolérable pour les parents – gavage par sonde naso-gastrique ou gastrostomie Problème de socialisation (ne peut manger qu'à la maison)

Tableau S9 (suite et fin)

Type de score	Valeur	Libellé
Score de comportement psycho-social	1	Pas de problème en général (ou petites difficultés très occasionnelles) Enfant de bonne humeur et très « facile » en général
	2	Inhibition ou opposition occasionnelles ou assez facilement résolues
	3	Inhibition ou opposition fréquentes et durables ; longues ou difficiles à apaiser perturbant souvent sur la vie familiale ou scolaire
	4	Inhibition ou opposition quotidienne, perturbant gravement la vie familiale et/ou altérant durablement la socialisation de l'enfant et/ou de sa famille
Score visuel	1	Bonne vision de chaque œil et vision des reliefs normale Fixation stable et poursuite oculaire normale sur chaque œil Si trouble de réfraction : modéré, ne nécessitant pas de correction Pas de gêne à l'occlusion
	2	Exophorie intermittente Trouble de réfraction nécessitant une correction Si amblyopie : corrigée par les lunettes Port de cache
	3	Trouble de réfraction sévère d'au moins 1 œil Amblyopie partiellement corrigée par les lunettes Strabisme manifeste Pas de vision binoculaire
	4	Amblyopie bilatérale Fixation absente ou très instable Poursuite oculaire absente ou très limitée
Score auditif	1	Aucune anomalie ; si explorations : normales
	2	Surdité unilatérale ou bilatérale légère (perte de 21 à 40 dB)
	3	Surdité bilatérale moyenne (perte de 41 à 70 dB)
	4	Surdité bilatérale sévère ou profonde (perte > 70 dB)
Appréciation globale des parents	1	Tout va pour le mieux (si problèmes : bien acceptés et bien gérés)
	2	Contraintes modérées / inquiétudes occasionnelles ayant globalement peu d'impact sur l'équilibre familial
	3	Inquiétude et contraintes importantes. Impact notable sur la famille
	4	Impact très lourd, ayant totalement bouleversé la famille

Annexe 4 : Analyses complémentaires

Tableau S10 : Facteurs associés à l'absence de visite à deux ans dans les régressions logistiques

univariées et multivariées avant imputation multiple, cohorte HYGIE-SEV, 1er novembre 2015-31

octobre 2022 (N=9 958)

Variable	Visite réalisée (N = 6 077)		Visite non réalisée (N = 3 881)		OR	IC 95%
	n	%	n	%		
Niveau d'éducation de la mère						
Supérieur	2 940	54,1	1 454	42,5	1,00	[Référence]
Secondaire (Collège/CAP/Lycée/BAC)	2 249	41,4	1 751	51,1	1,57	[1,44 ; 1,72]
Aucun ou niveau primaire	247	4,5	219	6,4	1,79	[1,48 ; 2,17]
Inconnu*	641	10,5	457	11,8	1,44	[1,26 ; 1,65]
Ressource économique du foyer						
Travail d'un parent au moins	4 851	88,8	2 829	82,7	1,00	[Référence]
Aides sociales	426	7,8	410	12,0	1,65	[1,43 ; 1,90]
Autre source	61	1,1	81	2,4	2,28	[1,63 ; 3,20]
Aucune	124	2,3	102	3,0	1,41	[1,08 ; 1,84]
Inconnu*	569	9,4	402	10,5	1,21	[1,06 ; 1,39]
Situation socio-professionnelle de la mère						
Travail : cadre et profession intellectuelle	1 063	17,5	452	11,6	1,00	[Référence]
Travail : intermédiaire	340	5,6	167	4,3	1,16	[0,93 ; 1,43]
Travail : ouvrier ou employé	1 695	27,9	960	24,7	1,33	[1,16 ; 1,53]
Travail : profession inconnue	164	2,7	87	2,2	1,25	[0,94 ; 1,65]
Congé parental	1 284	21,1	948	24,4	1,74	[1,51 ; 2,00]
Chômage/Sans emploi	1 097	18,1	926	23,9	1,99	[1,73 ; 2,29]
Inconnu*	434	7,1	341	8,8	1,85	[1,54 ; 2,21]
Situation socio-professionnelle du père						
Travail : cadre et profession intellectuelle	1 282	21,1	557	14,4	1,00	[Référence]
Travail : intermédiaire	663	10,9	411	10,6	1,43	[1,22 ; 1,67]
Travail : ouvrier ou employé	2 531	41,6	1 653	42,6	1,50	[1,34 ; 1,69]
Travail : profession inconnue	225	3,7	149	3,8	1,52	[1,21 ; 1,92]
Congé parental	42	0,7	40	1,0	2,19	[1,40 ; 3,42]
Chômage/Sans emploi	616	10,1	509	13,1	1,90	[1,63 ; 2,22]
Inconnu*	718	11,8	562	14,5	1,80	[1,55 ; 2,09]
Couverture sociale						
Sécurité sociale	4 621	86,5	2 636	79,2	1,00	[Référence]
Aucune ou AME ou CMU	722	13,5	693	20,8	1,68	[1,50 ; 1,89]
Inconnu*	734	12,1	552	14,2	1,32	[1,17 ; 1,49]
Conditions de logement						
Adéquates	4 068	78,1	2 345	72,7	1,00	[Référence]
Inadéquates (logement trop petit ou insalubre)	431	8,3	295	9,2	1,19	[1,01 ; 1,39]
Chez un proche	355	6,8	293	9,1	1,43	[1,22 ; 1,68]
Foyer d'accueil, établissement d'urgence ou autre	355	6,8	291	9,0	1,42	[1,21 ; 1,67]
Inconnu*	822	13,6	600	15,7	1,27	[1,13 ; 1,42]

Tableau S10 (suite)

Variable	Visite réalisée (N = 6 077)		Visite non réalisée (N = 3 881)		OR	IC 95%
	n	%	n	%		
Nombre de frères et sœurs au moment de la naissance						
0	1 620	34,1	847	28,5	1,00	[Référence]
1	1 560	32,8	934	31,4	1,15	[1,02 ; 1,29]
2 ou plus	1 572	33,1	1 190	40,1	1,45	[1,29 ; 1,62]
Inconnu*	1 281	21,2	861	22,5	1,29	[1,14 ; 1,45]
Statut marital de la mère						
En couple	4 968	88,4	3 016	84,8	1,00	[Référence]
Célibataire	655	11,6	542	15,2	1,36	[1,21 ; 1,54]
Inconnu*	454	7,5	323	8,3	1,17	[1,01 ; 1,36]
Sexe de l'enfant						
Féminin	2 889	47,5	1 829	47,1	1,00	[Référence]
Masculin	3 188	52,5	2 052	52,9	1,02	[0,94 ; 1,10]
Age gestationnel (SA)						
<=28	2 075	34,1	1 117	28,8	0,72	[0,65 ; 0,79]
29-30	1 669	27,5	1 015	26,2	0,81	[0,73 ; 0,90]
31-32	2 333	38,4	1 749	45,1	1,00	[Référence]
Z-score du périmètre crânien à la naissance						
<-1	641	18,4	379	16,7	1,00	[Référence]
[-1;0[1 219	35,0	755	33,3	1,05	[0,90 ; 1,22]
>0	1 623	46,6	1 130	49,9	1,18	[1,02 ; 1,37]
Inconnu*	2 594	42,7	1 617	41,7	1,05	[0,92 ; 1,22]
Z-score du poids à la naissance						
<-1	1 143	18,9	642	16,7	1,00	[Référence]
[-1;0[2 470	40,9	1 550	40,2	1,12	[1,00 ; 1,25]
>0	2 428	40,2	1 659	43,1	1,22	[1,08 ; 1,37]
Inconnu*	36	0,6	30	0,8	1,48	[0,90 ; 2,43]
Delta du z-score du poids						
<-1	3 646	79,3	2 232	78,6	1,00	[Référence]
[-1;0[727	15,8	479	16,9	1,08	[0,95 ; 1,22]
>0	227	4,9	129	4,5	0,93	[0,74 ; 1,16]
Inconnu*	1 477	24,3	1 041	26,8	1,15	[1,05 ; 1,27]
Prématurité provoquée						
Non	2 700	50,8	1 737	52,0	1,00	[Référence]
Oui	2 613	49,2	1 602	48,0	0,95	[0,87 ; 1,04]
Inconnu*	764	12,6	542	14,0	1,10	[0,97 ; 1,25]
Procréation médicalement assistée						
Non	4 009	83,8	2 637	88,3	1,00	[Référence]
Oui	773	16,2	351	11,7	0,69	[0,60 ; 0,79]
Inconnu*	1 295	21,3	893	23,0	1,05	[0,95 ; 1,16]
Naissance multiple						
Non	4 256	70,1	2 819	72,7	1,00	[Référence]
Oui	1 819	29,9	1 061	27,3	0,88	[0,81 ; 0,96]
Inconnu*	2	0,0	1	0,0	0,75	[0,04 ; 7,88]
Age de la mère à la naissance (années)						
<20	57	1,0	76	2,0	2,25	[1,59 ; 3,19]
20-29	1 818	30,4	1 335	34,9	1,24	[1,13 ; 1,35]
30-39	3 164	52,9	1 878	49,1	1,00	[Référence]
>=40	944	15,8	539	14,1	0,96	[0,85 ; 1,08]
Inconnu	94	1,5	53	1,4	0,95	[0,67 ; 1,33]

Tableau S10 (suite)

Variable	Visite réalisée (N = 6 077)		Visite non réalisée (N = 3 881)		OR	IC 95%
	n	%	n	%		
Mode d'accouchement						
Voie basse	2 085	35,5	1 446	38,9	1,00	[Référence]
Césarienne	3 781	64,5	2 271	61,1	0,87	[0,80 ; 0,94]
Inconnu*	211	3,5	164	4,2	1,12	[0,90 ; 1,39]
Tabac durant la grossesse						
Non	4 972	92,8	3 008	89,7	1,00	[Référence]
Oui	383	7,2	346	10,3	1,49	[1,28 ; 1,74]
Inconnu*	722	11,9	527	13,6	1,21	[1,07 ; 1,36]
Alcool durant la grossesse						
Non	5 290	99,2	3 270	99,0	1,00	[Référence]
Oui	40	0,8	33	1,0	1,33	[0,84 ; 2,12]
Inconnu*	747	12,3	578	14,9	1,25	[1,11 ; 1,41]
Cannabis durant la grossesse						
Non	5 270	99,4	3 249	99,0	1,00	[Référence]
Oui	30	0,6	32	1,0	1,73	[1,05 ; 2,86]
Inconnu*	777	12,8	600	15,5	1,25	[1,12 ; 1,41]
Autre produit toxique durant la grossesse						
Non	5 138	99,7	3 167	99,8	1,00	[Référence]
Oui	17	0,3	7	0,2	0,67	[0,26 ; 1,55]
Inconnu*	922	15,2	707	18,2	1,24	[1,12 ; 1,39]
Produit de substitution durant la grossesse						
Non	5 224	99,7	3 216	99,8	1,00	[Référence]
Oui	15	0,3	6	0,2	0,65	[0,23 ; 1,60]
Inconnu*	838	13,8	659	17,0	1,28	[1,14 ; 1,43]
Corticostéroïdes anténataux						
Non	728	13,5	559	16,4	1,00	[Référence]
Oui	4 677	86,5	2 842	83,6	0,79	[0,70 ; 0,89]
Inconnu*	672	11,1	480	12,4	0,93	[0,79 ; 1,09]
Chorioamniotite						
Non	4 348	87,4	2 733	87,4	1,00	[Référence]
Oui	629	12,6	393	12,6	0,99	[0,87 ; 1,14]
Inconnu*	1 100	18,1	755	19,5	1,09	[0,98 ; 1,21]
Résultats auditifs						
Normal	4 913	87,9	3 150	87,5	1,00	[Référence]
A contrôler	454	8,1	293	8,1	1,01	[0,86 ; 1,17]
Examen non fait	225	4,0	156	4,3	1,08	[0,88 ; 1,33]
Inconnu*	485	8,0	282	7,3	0,91	[0,78 ; 1,06]
Rétinopathie stade 2 ou plus						
Non	5 906	97,1	3 817	98,4	1,00	[Référence]
Oui	171	2,9	64	1,6	0,71	[0,52 ; 0,96]
Corticostéroïdes postnataux						
Non	4 202	84,7	2 780	90,7	1,00	[Référence]
Oui	758	15,3	284	9,3	0,57	[0,49 ; 0,65]
Inconnu*	1 117	18,4	817	21,1	1,11	[1,00 ; 1,22]
Traitement anti-virus respiratoire syncytial						
Non	2 690	44,3	1 869	48,2	1,00	[Référence]
Oui	3 379	55,7	2 008	51,8	0,86	[0,79 ; 0,93]
Inconnu*	8	0,1	4	0,1	0,72	[0,19 ; 2,29]

Tableau S10 (suite)

Variable	Visite réalisée (N = 6 077)		Visite non réalisée (N = 3 881)		OR	IC 95%
	n	%	n	%		
Dysplasie bronchopulmonaire						
Non	4 764	78,5	3 282	84,6	1,00	[Référence]
Oui	1 303	21,5	596	15,4	0,66	[0,60 ; 0,74]
Inconnu*	10	0,2	3	0,1	0,44	[0,10 ; 1,42]
Pathologie fœtale additionnelle						
Non	5 726	94,3	3 646	94,0	1,00	[Référence]
Oui	346	5,7	233	6,0	1,06	[0,89 ; 1,25]
Inconnu*	5	0,1	2	0,1	0,63	[0,09 ; 2,92]
Evaluation neurologique						
Pas de risque	376	6,2	291	7,5	1,00	[Référence]
Risques faibles	5 165	85,0	3 274	84,4	0,82	[0,70 ; 0,96]
Risques modérés	428	7,0	245	6,3	0,74	[0,59 ; 0,92]
Risques élevés	108	1,8	71	1,8	0,85	[0,61 ; 1,19]
Intervention chirurgicale à la naissance						
Non	4 487	89,3	2 866	91,0	1,00	[Référence]
Oui	536	10,7	282	9,0	0,82	[0,71 ; 0,96]
Inconnu*	1 054	17,3	733	18,9	1,09	[0,98 ; 1,21]
Allaitement prévu à la sortie						
Oui	2 351	47,3	1 599	50,9	1,00	[Référence]
Non	2 618	52,7	1 542	49,1	0,87	[0,79 ; 0,95]
Inconnu*	1 108	18,2	740	19,1	0,98	[0,88 ; 1,10]
Délai d'hospitalisation (jours)						
0-30	481	8,7	463	13,0	1,00	[Référence]
31-50	1 755	31,6	1 323	37,0	0,78	[0,67 ; 0,90]
51-70	1 496	26,9	877	24,6	0,61	[0,52 ; 0,71]
> 70	1 825	32,8	909	25,4	0,53	[0,46 ; 0,62]
Inconnu*	520	8,6	309	13,4	0,62	[0,51 ; 0,75]
Niveau urbain du lieu de vie (nombre d'habitants)						
Moyen ou faible (<50 000)	1 385	25,8	1 018	29,3	1,00	[Référence]
Elevé (>50 000)	3 986	74,2	2 457	70,7	0,84	[0,76 ; 0,92]
Inconnu*	706	11,6	406	10,5	0,78	[0,68 ; 0,91]
Réseau d'inclusion						
RPSOF – Région Sud-Ouest	1 808	29,8	868	22,4	1,00	[Référence]
RSPP – Paris	1 105	18,2	749	19,3	1,41	[1,25 ; 1,60]
NEF – Région Nord-Est	1 285	21,1	905	23,3	1,47	[1,30 ; 1,65]
RPVM – Région Sud-Est	660	10,9	313	8,1	0,99	[0,84 ; 1,15]
RPVO – Région Nord-Ouest	549	9,0	692	17,8	2,63	[2,29 ; 3,02]
MYP A – Région Ouest	670	11,0	354	9,1	1,10	[0,94 ; 1,28]
Année de naissance de l'enfant						
2015	817	13,4	498	12,8	1,00	[Référence]
2016	1 293	21,3	703	18,1	0,89	[0,77 ; 1,03]
2017	1 332	21,9	795	20,5	0,98	[0,85 ; 1,13]
2018	1 152	19,0	850	21,9	1,21	[1,05 ; 1,40]
2019-2020	1 483	24,4	1 035	26,7	1,14	[1,00 ; 1,31]
Médecin pilote						
Hospitalier (public/privé)	3 418	60,4	2 522	70,4	1,00	[Référence]
Libéral	1 692	29,9	778	21,7	0,62	[0,56 ; 0,69]
Autre	547	9,7	284	7,9	0,70	[0,60 ; 0,82]
Inconnu*	420	6,9	297	7,7	0,96	[0,82 ; 1,12]

Tableau S10 (suite et fin)

Variable	Visite réalisée (N = 6 077)		Visite non réalisée (N = 3 881)		OR	IC 95%
	n	%	n	%		
Type de maternité						
Niveau 3	4 745	81,9	2 836	77,6	1,00	[Référence]
Niveau 1 ou 2	1 051	18,1	819	22,4	1,30	[1,18 ; 1,44]
Inconnu*	281	4,6	226	5,8	1,35	[1,12 ; 1,61]
Enfant né en dehors de son établissement d'inclusion						
Non	3 594	61,3	2 179	58,4	1,00	[Référence]
Oui	2 273	38,7	1 551	41,6	1,13	[1,04 ; 1,22]
Inconnu*	210	3,5	151	3,9	1,19	[0,95 ; 1,47]
Distance entre le lieu de domicile et le médecin pilote (km)						
<3	2 934	53,2	1 982	56,4	1,00	[Référence]
3-10	1 191	21,6	746	21,2	0,93	[0,83 ; 1,03]
>10	1 393	25,2	784	22,3	0,83	[0,75 ; 0,92]
Inconnu*	559	9,2	369	9,5	0,98	[0,85 ; 1,13]

Abréviations : IC, Intervalle de confiance ; OR, Odds Ratio ; SA, semaines d'aménorrhée

* Les proportions de données inconnues sont calculées par rapport au nombre total de sujets, tandis que les proportions dans chaque catégorie d'une variable donnée sont calculées par rapport au nombre de sujets ayant des données complètes pour cette variable.

Tableau S11 : Facteurs associés à l'absence de visite à cinq ans dans les régressions logistiques univariées et multivariées avant imputation multiple, cohorte HYGIE-SEV, 1er novembre 2015-31 octobre 2022 (N=3 898)

Variable	Visite réalisée (N = 1 458)		Visite non réalisée (N = 2 440)		OR	IC 95%
	n	%	n	%		
Niveau d'éducation de la mère						
Supérieur	762	58,2	1 004	45,6	1,00	[Référence]
Secondaire (Collège/CAP/Lycée/BAC)	491	37,5	1 078	49,0	1,67	[1,45 ; 1,92]
Aucun ou niveau primaire	57	4,4	119	5,4	1,58	[1,15 ; 2,22]
Inconnu*	148	10,2	239	9,8	1,23	[0,98 ; 1,54]
Ressource économique du foyer						
Travail d'un parent au moins	1 205	90,7	1 843	84,3	1,00	[Référence]
Aides sociales	78	5,9	235	10,8	1,97	[1,52 ; 2,58]
Autre source	21	1,6	44	2,0	1,37	[0,82 ; 2,36]
Aucune	25	1,9	64	2,9	1,67	[1,06 ; 2,72]
Inconnu*	118	8,2	231	9,6	1,28	[1,02 ; 1,62]
Situation socio-professionnelle de la mère						
Travail : cadre et profession intellectuelle	271	18,6	309	12,7	1,00	[Référence]
Travail : intermédiaire	88	6,0	125	5,1	1,25	[0,91 ; 1,71]
Travail : ouvrier ou employé	439	30,1	636	26,1	1,27	[1,04 ; 1,56]
Travail : profession inconnue	51	3,5	81	3,3	1,39	[0,95 ; 2,06]
Congé parental	271	18,6	628	25,7	2,03	[1,64 ; 2,52]
Chômage/Sans emploi	241	16,5	497	20,4	1,81	[1,45 ; 2,26]
Inconnu*	97	6,7	164	6,7	1,48	[1,10 ; 2,00]
Situation socio-professionnelle du père						
Travail : cadre et profession intellectuelle	341	23,4	394	16,1	1,00	[Référence]
Travail : intermédiaire	130	8,9	266	10,9	1,77	[1,37 ; 2,29]
Travail : ouvrier ou employé	582	39,9	1 047	42,9	1,56	[1,30 ; 1,86]
Travail : profession inconnue	85	5,8	113	4,6	1,15	[0,84 ; 1,58]
Congé parental	7	0,5	25	1,0	3,09	[1,39 ; 7,83]
Chômage/Sans emploi	148	10,2	293	12,0	1,71	[1,34 ; 2,19]
Inconnu*	165	11,3	302	12,4	1,58	[1,25 ; 2,01]
Couverture sociale						
Sécurité sociale	1 061	87,7	1 709	82,5	1,00	[Référence]
Aucune ou AME ou CMU	149	12,3	362	17,5	1,51	[1,23 ; 1,86]
Inconnu*	248	17,0	369	15,1	0,92	[0,77 ; 1,10]
Conditions de logement						
Adéquates	1 007	79,7	1 571	74,6	1,00	[Référence]
Inadéquates (logement trop petit ou insalubre)	116	9,2	188	8,9	1,04	[0,81 ; 1,33]
Chez un proche	75	5,9	187	8,9	1,60	[1,21 ; 2,12]
Foyer d'accueil, établissement d'urgence ou autre	65	5,1	161	7,6	1,59	[1,18 ; 2,15]
Inconnu*	184	12,7	310	12,8	1,08	[0,89 ; 1,32]
Nombre de frères et sœurs au moment de la naissance						
0	397	37,9	595	33,0	1,00	[Référence]
1	349	33,3	561	31,1	1,07	[0,89 ; 1,29]
2 ou plus	301	28,7	649	36,0	1,44	[1,19 ; 1,73]
Inconnu*	401	27,7	613	25,4	1,02	[0,85 ; 1,22]

Tableau S11 (suite)

Variable	Visite réalisée (N = 1 458)		Visite non réalisée (N = 2 440)		OR	IC 95%
	n	%	n	%		
Statut marital de la mère						
En couple	1 213	88,5	1 978	86,8	1,00	[Référence]
Célibataire	158	11,5	302	13,2	1,17	[0,96 ; 1,44]
Inconnu*	87	6,0	160	6,6	1,13	[0,86 ; 1,48]
Sexe de l'enfant						
Féminin	687	47,1	1 122	46,0	1,00	[Référence]
Masculin	771	52,9	1 318	54,0	1,05	[0,92 ; 1,19]
Age gestationnel (SA)						
<=28	514	35,3	748	30,7	0,73	[0,63 ; 0,85]
29-30	416	28,5	644	26,4	0,78	[0,66 ; 0,92]
31-32	528	36,2	1 048	43,0	1,00	[Référence]
Z-score du périmètre crânien à la naissance						
<-1	166	20,0	243	17,2	1,00	[Référence]
[-1;0[277	33,5	470	33,2	1,16	[0,90 ; 1,48]
>0	385	46,5	702	49,6	1,25	[0,99 ; 1,57]
Inconnu*	630	43,2	1 025	42,0	1,11	[0,89 ; 1,39]
Z-score du poids à la naissance						
<-1	294	20,3	440	18,1	1,00	[Référence]
[-1;0[585	40,4	983	40,4	1,12	[0,94 ; 1,34]
>0	570	39,3	1 012	41,6	1,19	[0,99 ; 1,42]
Inconnu*	9	0,6	5	0,2	0,37	[0,11 ; 1,09]
Delta du z-score du poids						
<-1	806	83,6	1 319	81,2	1,00	[Référence]
[-1;0[124	12,9	256	15,8	1,26	[1,00 ; 1,59]
>0	34	3,5	50	3,1	0,90	[0,58 ; 1,41]
Inconnu*	494	33,9	815	33,4	1,01	[0,87 ; 1,16]
Prématurité provoquée						
Non	580	47,8	1 048	50,2	1,00	[Référence]
Oui	633	52,2	1 038	49,8	0,91	[0,79 ; 1,05]
Inconnu*	245	16,8	354	14,5	0,80	[0,66 ; 0,97]
Procréation médicalement assistée						
Non	913	83,6	1 624	86,2	1,00	[Référence]
Oui	179	16,4	259	13,8	0,81	[0,66 ; 1,00]
Inconnu*	366	25,1	557	22,8	0,86	[0,73 ; 1,00]
Naissance multiple						
Non	999	68,5	1 669	68,4	1,00	[Référence]
Oui	459	31,5	771	31,6	1,01	[0,87 ; 1,16]
Age de la mère à la naissance (années)						
<20	12	0,8	50	2,1	2,90	[1,59 ; 5,74]
20-29	403	28,1	858	35,7	1,48	[1,28 ; 1,72]
30-39	785	54,8	1 127	46,9	1,00	[Référence]
>=40	232	16,2	370	15,4	1,11	[0,92 ; 1,34]
Inconnu	26	1,8	35	1,4	0,94	[0,56 ; 1,58]
Mode d'accouchement						
Voie basse	460	33,2	829	35,6	1,00	[Référence]
Césarienne	925	66,8	1 499	64,4	0,90	[0,78 ; 1,03]
Inconnu*	73	5,0	112	4,6	0,85	[0,62 ; 1,17]

Tableau S11 (suite)

Variable	Visite réalisée (N = 1 458)		Visite non réalisée (N = 2 440)		OR	IC 95%
	n	%	n	%		
Tabac durant la grossesse						
Non	1 235	92,6	1 966	89,1	1,00	[Référence]
Oui	98	7,4	240	10,9	1,54	[1,21 ; 1,97]
Inconnu*	125	8,6	234	9,6	1,18	[0,94 ; 1,48]
Alcool durant la grossesse						
Non	1 320	99,2	2 157	98,8	1,00	[Référence]
Oui	11	0,8	26	1,2	1,45	[0,73 ; 3,06]
Inconnu*	127	8,7	257	10,5	1,24	[0,99 ; 1,55]
Cannabis durant la grossesse						
Non	1 303	99,6	2 144	99,3	1,00	[Référence]
Oui	5	0,4	15	0,7	1,82	[0,70 ; 5,62]
Inconnu*	150	10,3	281	11,5	1,14	[0,92 ; 1,41]
Autre produit toxique durant la grossesse						
Non	1 269	99,6	2 076	99,5	1,00	[Référence]
Oui	5	0,4	11	0,5	1,34	[0,49 ; 4,28]
Inconnu*	184	12,6	353	14,5	1,17	[0,97 ; 1,42]
Produit de substitution durant la grossesse						
Non	1 294	99,5	2 113	99,7	1,00	[Référence]
Oui	6	0,5	7	0,3	0,71	[0,24 ; 2,22]
Inconnu*	158	10,8	320	13,1	1,24	[1,01 ; 1,52]
Corticostéroïdes anténataux						
Non	161	13,0	342	16,2	1,00	[Référence]
Oui	1 078	87,0	1 772	83,8	0,77	[0,63 ; 0,95]
Inconnu*	219	15,0	326	13,4	0,70	[0,54 ; 0,90]
Chorioamniotite						
Non	972	86,2	1 708	87,5	1,00	[Référence]
Oui	156	13,8	244	12,5	0,89	[0,72 ; 1,11]
Inconnu*	330	22,6	488	20,0	0,84	[0,72 ; 0,99]
Résultats auditifs						
Normal	1 143	86,5	1 984	87,7	1,00	[Référence]
A contrôler	124	9,4	186	8,2	0,86	[0,68 ; 1,10]
Examen non fait	54	4,1	91	4,0	0,97	[0,69 ; 1,38]
Inconnu*	137	9,4	179	7,3	0,75	[0,60 ; 0,95]
Rétinopathie stade 2 ou plus						
Non	1 422	97,5	2 389	97,9	1,00	[Référence]
Oui	36	2,5	51	2,1	0,91	[0,58 ; 1,45]
Corticostéroïdes postnataux						
Non	934	81,9	1 683	88,0	1,00	[Référence]
Oui	207	18,1	229	12,0	0,61	[0,50 ; 0,75]
Inconnu*	317	21,7	528	21,6	0,92	[0,79 ; 1,09]
Traitement anti-virus respiratoire syncytial						
Non	611	42,0	1 153	47,3	1,00	[Référence]
Oui	845	58,0	1 285	52,7	0,81	[0,71 ; 0,92]
Inconnu*	2	0,1	2	0,1	0,53	[0,06 ; 4,43]
Dysplasie bronchopulmonaire						
Non	1 121	77,1	1 997	81,9	1,00	[Référence]
Oui	333	22,9	441	18,1	0,74	[0,63 ; 0,87]
Inconnu*	4	0,3	2	0,1	0,28	[0,04 ; 1,44]
Pathologie fœtale additionnelle						
Non	1 363	93,5	2 286	93,7	1,00	[Référence]
Oui	95	6,5	154	6,3	0,97	[0,74 ; 1,26]
Inconnu*	1 320	99,2	2 157	98,8		

Tableau S11 (suite et fin)

Variable	Visite réalisée (N = 1 458)		Visite non réalisée (N = 2 440)		OR	IC 95%
	n	%	n	%		
Evaluation neurologique						
Pas de risque	80	5,5	164	6,7	1,00	[Référence]
Risques faibles	1 244	85,3	2 040	83,6	0,80	[0,60 ; 1,05]
Risques modérés	111	7,6	191	7,8	0,84	[0,59 ; 1,20]
Risques élevés	23	1,6	45	1,8	0,95	[0,54 ; 1,71]
Intervention chirurgicale à la naissance						
Non	1 052	89,8	1 843	90,5	1,00	[Référence]
Oui	120	10,2	193	9,5	0,92	[0,72 ; 1,17]
Inconnu*	286	19,6	404	16,6	0,81	[0,68 ; 0,96]
Allaitement prévu à la sortie						
Oui	613	50,1	1 010	50,3	1,00	[Référence]
Non	611	49,9	999	49,7	0,99	[0,86 ; 1,14]
Inconnu*	234	16,0	431	17,7	1,12	[0,93 ; 1,35]
Délai d'hospitalisation (jours)						
0-30	99	7,4	258	11,5	1,00	[Référence]
31-50	421	31,6	797	35,6	0,74	[0,57 ; 0,95]
51-70	338	25,4	553	24,7	0,67	[0,51 ; 0,87]
> 70	475	35,6	628	28,1	0,55	[0,42 ; 0,71]
Inconnu*	125	8,4	204	8,4	0,67	[0,49 ; 0,92]
Niveau urbain du lieu de vie (nombre d'habitants)						
Moyen ou faible (<50 000)	354	28,3	650	30,6	1,00	[Référence]
Elevé (> 50 000)	898	71,7	1 477	69,4	0,90	[0,77 ; 1,04]
Inconnu*	206	14,1	313	12,8	0,83	[0,67 ; 1,03]
Réseau d'inclusion						
RPSOF – Région Sud-Ouest	397	27,2	579	23,7	1,00	[Référence]
RSPP – Paris	248	17,0	455	18,6	1,26	[1,03 ; 1,54]
NEF – Région Nord-Est	339	23,3	485	19,9	0,98	[0,81 ; 1,18]
RPVM – Région Sud-Est	137	9,4	214	8,8	1,07	[0,84 ; 1,38]
RPVO – Région Nord-Ouest	96	6,6	406	16,6	2,90	[2,25 ; 3,76]
MYPA – Région Ouest	241	16,5	301	12,3	0,86	[0,69 ; 1,06]
Année de naissance de l'enfant						
2015	555	38,1	746	30,6	1,00	[Référence]
2016-2017	903	61,9	1 694	69,4	1,40	[1,22 ; 1,60]
Médecin pilote						
Hospitalier (public/privé)	799	59,6	1 538	68,6	1,00	[Référence]
Libéral	425	31,7	518	23,1	0,63	[0,54 ; 0,74]
Autre	116	8,7	185	8,3	0,83	[0,65 ; 1,06]
Inconnu*	118	8,1	199	8,2	0,88	[0,69 ; 1,12]
Type de maternité						
Niveau 3	1 138	83,6	1 839	79,8	1,00	[Référence]
Niveau 1 ou 2	224	16,4	465	20,2	1,28	[1,08 ; 1,53]
Inconnu*	96	6,6	136	5,6	0,88	[0,67 ; 1,15]
Enfant né en dehors de son établissement d'inclusion						
Non	899	65,1	1 477	63,0	1,00	[Référence]
Oui	481	34,9	869	37,0	1,10	[0,96 ; 1,26]
Inconnu*	78	5,3	94	3,9	0,73	[0,54 ; 1,00]
Distance entre le lieu de domicile et le médecin pilote (km)						
<3	694	52,1	1 212	55,6	1,00	[Référence]
3-10	297	22,3	479	22,0	0,92	[0,78 ; 1,10]
>10	340	25,5	490	22,5	0,83	[0,70 ; 0,98]
Inconnu*	127	8,7	259	10,6	1,17	[0,93 ; 1,48]

Abréviations : IC, Intervalle de confiance ; OR, Odds Ratio ; SA, semaines d'aménorrhée

* Les proportions de données inconnues sont calculées par rapport au nombre total de sujets, tandis que les proportions dans chaque catégorie d'une variable donnée sont calculées par rapport au nombre de sujets ayant des données complètes pour cette variable.

Tableau S12 : Facteurs associés à la perte de vue à deux ans (définition plus stricte) dans les régressions logistiques univariées et multivariées après imputation multiple, cohorte HYGIE-SEV, 1er novembre 2015-31 octobre 2022 (N=3 898)

Variable	Univarié			Multivarié	
	OR	IC 95%	P	OR	IC 95%
Niveau d'éducation de la mère			<0,001		
Supérieur	1,00	[Référence]		1,00	[Référence]
Secondaire (Collège/CAP/Lycée/BAC)	1,75	[1,51 ; 2,02]		1,34	[1,11 ; 1,61]
Aucun ou niveau primaire	1,39	[1,00 ; 1,93]		1,00	[0,68 ; 1,47]
Ressource économique du foyer			<0,001		
Travail d'un parent au moins	1,00	[Référence]		1,00	[Référence]
Aides sociales	1,62	[1,28 ; 2,03]		0,95	[0,68 ; 1,32]
Autre source	1,96	[1,22 ; 3,12]		1,33	[0,78 ; 2,27]
Aucune	1,50	[0,99 ; 2,27]		0,87	[0,50 ; 1,51]
Situation socio-professionnelle de la mère			<0,001		
Travail : cadre et profession intellectuelle	1,00	[Référence]		1,00	[Référence]
Travail : intermédiaire	1,37	[0,97 ; 1,95]		1,03	[0,70 ; 1,50]
Travail : ouvrier ou employé	1,34	[1,06 ; 1,70]		0,93	[0,71 ; 1,23]
Congé parental	1,93	[1,53 ; 2,45]		1,34	[1,00 ; 1,79]
Chômage/Sans emploi	2,05	[1,61 ; 2,61]		1,22	[0,90 ; 1,67]
Situation socio-professionnelle du père			<0,001		
Travail : cadre et profession intellectuelle	1,00	[Référence]		1,00	[Référence]
Travail : intermédiaire	1,70	[1,31 ; 2,22]		1,49	[1,11 ; 1,99]
Travail : ouvrier ou employé	1,51	[1,24 ; 1,84]		1,12	[0,88 ; 1,43]
Congé parental	2,62	[1,42 ; 4,83]		1,55	[0,79 ; 3,04]
Chômage/Sans emploi	2,11	[1,67 ; 2,66]		1,36	[0,99 ; 1,87]
Couverture sociale			<0,001		
Sécurité sociale	1,00	[Référence]		1,00	[Référence]
Aucune ou AME ou CMU	1,54	[1,22 ; 1,94]		1,10	[0,84 ; 1,44]
Conditions de logement			0,002		
Adéquates	1,00	[Référence]		1,00	[Référence]
Inadéquates (logement trop petit ou insalubre)	1,08	[0,83 ; 1,42]		0,94	[0,69 ; 1,27]
Chez un proche	1,59	[1,22 ; 2,06]		1,20	[0,88 ; 1,65]
Foyer d'accueil, établissement d'urgence ou autre	1,73	[1,27 ; 2,35]		1,50	[1,05 ; 2,14]
Nombre de frères et sœurs à la naissance			0,026		
0	1,00	[Référence]		1,00	[Référence]
1	1,13	[0,93 ; 1,36]		1,10	[0,90 ; 1,35]
2 ou plus	1,30	[1,07 ; 1,57]		1,23	[0,99 ; 1,54]
Statut marital de la mère			0,008		
En couple	1,00	[Référence]		1,00	[Référence]
Célibataire	1,31	[1,07 ; 1,60]		0,94	[0,73 ; 1,20]
Sexe de l'enfant			0,378		
Féminin	1,00	[Référence]		1,00	[Référence]
Masculin	1,06	[0,93 ; 1,22]		1,04	[0,90 ; 1,20]
Age gestationnel (SA)			<0,001		
<=28	0,64	[0,54 ; 0,75]		0,70	[0,56 ; 0,87]
29-30	0,72	[0,61 ; 0,85]		0,75	[0,62 ; 0,90]
31-32	1,00	[Référence]		1,00	[Référence]

Tableau S12 (suite)

Variable	Univarié			Multivarié	
	OR	IC 95%	P	OR	IC 95%
Z-score du périmètre crânien à la naissance			0,268		
<-1	1,00	[Référence]			
[-1;0[1,07	[0,85 ; 1,35]			
>0	1,17	[0,95 ; 1,45]			
Z-score du poids à la naissance			0,238		
<-1	1,00	[Référence]			
[-1;0[1,18	[0,94 ; 1,43]			
>0	1,19	[0,98 ; 1,44]			
Delta du z-score du poids			0,639		
<-1	1,00	[Référence]			
[-1;0[1,10	[0,89 ; 1,37]			
>0	0,97	[0,62 ; 1,52]			
Prématurité provoquée			0,472		
Non	1,00	[Référence]			
Oui	0,95	[0,82 ; 1,09]			
Procréation médicalement assistée			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,62	[0,49 ; 0,77]		0,75	[0,58 ; 0,96]
Naissance multiple			0,810		
Non	1,00	[Référence]			
Oui	1,02	[0,88 ; 1,17]			
Age de la mère à la naissance (années)			<0,001		
<20	2,20	[1,32 ; 3,65]		1,40	[0,80 ; 2,46]
20-29	1,33	[1,14 ; 1,55]		1,19	[1,00 ; 1,42]
30-39	1,00	[Référence]		1,00	[Référence]
>=40	0,95	[0,77 ; 1,16]		0,94	[0,75 ; 1,17]
Mode d'accouchement			0,850		
Voie basse	1,00	[Référence]			
Césarienne	1,01	[0,87 ; 1,17]			
Tabac durant la grossesse			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	1,63	[1,30 ; 2,03]		1,40	[1,09 ; 1,81]
Alcool durant la grossesse			0,419		
Non	1,00	[Référence]			
Oui	1,33	[0,68 ; 2,58]			
Cannabis durant la grossesse			0,175		
Non	1,00	[Référence]		1,00	[Référence]
Oui	1,62	[0,81 ; 3,22]		1,08	[0,49 ; 2,37]
Autre produit toxique durant la grossesse			0,490		
Non	1,00	[Référence]			
Oui	1,16	[0,54 ; 2,48]			
Produit de substitution durant la grossesse			0,500		
Non	1,00	[Référence]			
Oui	0,88	[0,35 ; 2,25]			
Corticostéroïdes anténataux			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,69	[0,56 ; 0,85]		0,79	[0,64 ; 0,99]
Chorioamniotite			0,654		
Non	1,00	[Référence]			
Oui	0,99	[0,78 ; 1,24]			

Tableau S12 (suite)

Variable	Univarié			Multivarié	
	OR	IC 95%	P	OR	IC 95%
Résultats auditifs			0,108		
Normal	1,00	[Référence]			
A contrôler	0,79	[0,61 ; 1,03]			
Examen non fait	0,77	[0,53 ; 1,14]			
Rétinopathie stade 2 ou plus			0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,37	[0,21 ; 0,68]		0,56	[0,30 ; 1,05]
Corticostéroïdes postnataux			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,65	[0,52 ; 0,82]		0,93	[0,69 ; 1,25]
Traitement anti-virus respiratoire syncytial			0,003		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,82	[0,71 ; 0,93]		0,89	[0,75 ; 1,06]
Dysplasie bronchopulmonaire			<0,001		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,67	[0,56 ; 0,81]		0,85	[0,66 ; 1,08]
Pathologie fœtale additionnelle			0,602		
Non	1,00	[Référence]			
Oui	0,93	[0,70 ; 1,23]			
Evaluation neurologique			0,096		
Pas de risque	1,00	[Référence]		1,00	[Référence]
Risques faibles	0,74	[0,58 ; 0,96]		0,82	[0,61 ; 1,10]
Risques modérés	0,69	[0,48 ; 0,99]		0,85	[0,58 ; 1,26]
Risques élevés	0,94	[0,54 ; 1,65]		1,02	[0,56 ; 1,85]
Intervention chirurgicale à la naissance			0,084		
Non	1,00	[Référence]		1,00	[Référence]
Oui	0,80	[0,62 ; 1,03]		0,94	[0,72 ; 1,24]
Allaitement prévu à la sortie			0,043		
Oui	1,00	[Référence]		1,00	[Référence]
Non	0,86	[0,75 ; 0,99]		0,91	[0,78 ; 1,06]
Délai d'hospitalisation (jours)			<0,001		
0-30	1,00	[Référence]			
31-50	0,67	[0,52 ; 0,85]			
51-70	0,56	[0,44 ; 0,72]			
> 70	0,44	[0,35 ; 0,57]			
Niveau urbain du lieu de vie (nombre d'habitants)			0,001		
Moyen ou faible (<50 000)	1,00	[Référence]		1,00	[Référence]
Elevé (>50 000)	0,77	[0,66 ; 0,90]		0,94	[0,76 ; 1,16]
Réseau d'inclusion			<0,001		
RPSOF – Région Sud-Ouest	1,00	[Référence]		1,00	[Référence]
RSPP – Paris	1,06	[0,86 ; 1,32]		1,22	[0,97 ; 1,55]
NEF – Région Nord-Est	1,32	[1,08 ; 1,61]		1,05	[0,84 ; 1,31]
RPVM – Région Sud-Est	0,77	[0,58 ; 1,03]		0,89	[0,65 ; 1,21]
RPVO – Région Nord-Ouest	2,36	[1,88 ; 2,95]		2,35	[1,80 ; 3,06]
MYPA – Région Ouest	1,14	[0,91 ; 1,44]		1,35	[1,00 ; 1,82]
Année de naissance de l'enfant			0,873		
2015	1,00	[Référence]		1,00	[Référence]
2016-2017	1,01	[0,88 ; 1,17]		1,00	[0,85 ; 1,18]

Tableau S12 (suite et fin)

Variable	Univarié			Multivarié	
	OR	IC 95%	P	OR	IC 95%
Médecin pilote			<0,001		
Hospitalier (public/privé)	1,00	[Référence]		1,00	[Référence]
Libéral	0,76	[0,65 ; 0,89]		0,87	[0,72 ; 1,05]
Autre	0,67	[0,51 ; 0,88]		0,57	[0,42 ; 0,77]
Type de maternité			<0,001		
Niveau 3	1,00	[Référence]		1,00	[Référence]
Niveau 1 ou 2	1,53	[1,29 ; 1,81]		1,20	[0,98 ; 1,47]
Enfant né en dehors de son établissement d'inclusion			0,052		
Non	1,00	[Référence]		1,00	[Référence]
Oui	1,18	[1,02 ; 1,36]		1,22	[1,04 ; 1,42]
Distance entre le lieu de domicile et le médecin pilote (km)			0,269		
<3	1,00	[Référence]			
3-10	0,98	[0,82 ; 1,16]			
>10	0,84	[0,71 ; 1,01]			

Abréviations : IC, Intervalle de confiance ; OR, Odds Ratio ; SA, semaines d'aménorrhée