Identification de maladies associées aux régressions du développement dans le syndrome de Phelan-McDermid : analyse exhaustive d'un registre international

Mikaël Dusenne

UFR DE SANTE DE ROUEN NORMANDIE

ANNEE 2019 **N°**

THESE POUR LE
DOCTORAT EN MEDECINE

(Diplôme d'État)

Par
Mikaël Dusenne
NE LE 08/08/1988 A Lille

PRESENTEE ET SOUTENUE PUBLIQUEMENT LE 29/11/2019

IDENTIFICATION DE MALADIES ASSOCIÉES AUX RÉGRESSIONS DU DÉVELOPPEMENT DANS LE SYNDROME DE PHELAN-MCDERMID: ANALYSE EXHAUSTIVE D'UN REGISTRE INTERNATIONAL

PRÉSIDENT DE JURY:
Monsieur le Professeur **Stéfan Darmoni**
DIRECTEUR DE THÈSE:
Monsieur le Professeur **Stéfan Darmoni**
MEMBRES DU JURY:
Monsieur le Professeur **Ygal Benhamou**
Madame le Maître de conférence **Laetitia Rollin**
Monsieur le Docteur **Jean-Philippe Leroy**

ANNEE UNIVERSITAIRE 2018 - 2019

U.F.R. SANTÉ DE ROUEN

DOYEN : **Professeur Benoît VEBER**

ASSESSEURS : **Professeur Michel GUERBET**

Professeur Agnès LIARD-ZMUDA

Professeur Guillaume SAVOYE

I - **MEDECINE**

PROFESSEURS DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mr Frédéric **ANSELME**	HCN	Cardiologie
Mme Gisèle **APTER**	Havre	Pédopsychiatrie
Mme Isabelle **AUQUIT AUCKBUR**	HCN	Chirurgie plastique
Mr Jean-Marc **BASTE**	HCN	Chirurgie Thoracique
Mr Fabrice **BAUER**	HCN	Cardiologie
Mme Soumeya **BEKRI**	HCN	Biochimie et biologie moléculaire
Mr Ygal **BENHAMOU**	HCN	Médecine interne
Mr Jacques **BENICHOU**	HCN	Bio statistiques et informatique médicale
Mr Olivier **BOYER**	UFR	Immunologie
Mme Sophie **CANDON**	HCN	Immunologie
Mr François **CARON**	HCN	Maladies infectieuses et tropicales
Mr Philippe **CHASSAGNE**	HCN	Médecine interne (gériatrie)
Mr Vincent **COMPERE**	HCN	Anesthésiologie et réanimation chirurgicale
Mr Jean-Nicolas **CORNU**	HCN	Urologie
Mr Antoine **CUVELIER**	HB	Pneumologie
Mr Jean-Nicolas **DACHER**	HCN	Radiologie et imagerie médicale
Mr Stéfan **DARMONI**	HCN	Informatique médicale et techniques de communication
Mr Pierre **DECHELOTTE**	HCN	Nutrition
Mr Stéphane **DERREY**	HCN	Neurochirurgie
Mr Frédéric **DI FIORE**	HCH-CB	Cancérologie

Mr Fabien **DOGUET**	HCN	Chirurgie Cardio Vasculaire
Mr Jean **DOUCET**	SJ	Thérapeutique - Médecine interne et gériatrie
Mr Bernard **DUBRAY**	CB	Radiothérapie
Mr Frank **DUJARDIN**	HCN	Chirurgie orthopédique - Traumatologique
Mr Fabrice **DUPARC**	HCN	Anatomie - Chirurgie orthopédique et traumatologique
Mr Eric **DURAND**	HCN	Cardiologie
Mr Bertrand **DUREUIL**	HCN	Anesthésiologie et réanimation chirurgicale
Mme Hélène **ELTCHANINOFF**	HCN	Cardiologie
Mr Manuel **ETIENNE**	HCN	Maladies infectieuses et tropicales
Mr Thierry **FREBOURG**	UFR	Génétique
Mr Pierre **FREGER**	HCN	Anatomie - Neurochirurgie
Mr Jean François **GEHANNO**	HCN	Médecine et santé au travail
Mr Emmanuel **GERARDIN**	HCN	Imagerie médicale
Mme Priscille **GERARDIN**	HCN	Pédopsychiatrie
M. Guillaume **GOURCEROL**	HCN	Physiologie
Mr Dominique **GUERROT**	HCN	Néphrologie
Mr Olivier **GUILLIN**	HCN	Psychiatrie Adultes
Mr Didier **HANNEQUIN**	HCN	Neurologie
Mr Claude **HOUDAYER**	HCN	Génétique
Mr Fabrice **JARDIN**	CB	Hématologie
Mr Luc-Marie **JOLY**	HCN	Médecine d'urgence
Mr Pascal **JOLY**	HCN	Dermato – Vénéréologie
Mme Bouchra **LAMIA**	Havre	Pneumologie
Mme Annie **LAQUERRIERE**	HCN	Anatomie et cytologie pathologiques
Mr Vincent **LAUDENBACH**	HCN	Anesthésie et réanimation chirurgicale
Mr Joël **LECHEVALLIER**	HCN	Chirurgie infantile
Mr Hervé **LEFEBVRE**	HB	Endocrinologie et maladies métaboliques
Mr Thierry **LEQUERRE**	HB	Rhumatologie
Mme Anne-Marie **LEROI**	HCN	Physiologie
Mr Hervé **LEVESQUE**	HB	Médecine interne
Mme Agnès **LIARD-ZMUDA**	HCN	Chirurgie Infantile
Mr Pierre Yves **LITZLER**	HCN	Chirurgie cardiaque
Mr Bertrand **MACE**	HCN	Histologie, embryologie, cytogénétique
M. David **MALTETE**	HCN	Neurologie
Mr Christophe **MARGUET**	HCN	Pédiatrie
Mme Isabelle **MARIE**	HB	Médecine interne
Mr Jean-Paul **MARIE**	HCN	Oto-rhino-laryngologie
Mr Loïc **MARPEAU**	HCN	Gynécologie - Obstétrique

Mr Stéphane **MARRET**	HCN	Pédiatrie
Mme Véronique **MERLE**	HCN	Epidémiologie
Mr Pierre **MICHEL**	HCN	Hépato-gastro-entérologie
M. Benoit **MISSET** *(détachement)*	HCN	Réanimation Médicale Mr Jean-François
M. **Jean-François MUIR** *(surnombre)*	HB	Pneumologie
Mr Marc **MURAINE**	HCN	Ophtalmologie
Mr Christophe **PEILLON**	HCN	Chirurgie générale
Mr Christian **PFISTER**	HCN	Urologie
Mr Jean-Christophe **PLANTIER**	HCN	Bactériologie - Virologie
Mr Didier **PLISSONNIER**	HCN	Chirurgie vasculaire
Mr Gaëtan **PREVOST**	HCN	Endocrinologie
Mr Jean-Christophe **RICHARD** *(détachement)*	HCN	Réanimation médicale - Médecine d'urgence
Mr Vincent **RICHARD**	UFR	Pharmacologie
Mme Nathalie **RIVES**	HCN	Biologie du développement et de la reproduction
Mr Horace **ROMAN** *(disponibilité)*	HCN	Gynécologie - Obstétrique
Mr Jean-Christophe **SABOURIN**	HCN	Anatomie - Pathologie
Mr Guillaume **SAVOYE**	HCN	Hépato-gastrologie
Mme Céline **SAVOYE–COLLET**	HCN	Imagerie médicale
Mme Pascale **SCHNEIDER**	HCN	Pédiatrie
Mr Michel **SCOTTE**	HCN	Chirurgie digestive
Mme Fabienne **TAMION**	HCN	Thérapeutique
Mr Luc **THIBERVILLE**	HCN	Pneumologie
Mr Christian **THUILLEZ** *(surnombre)*	HB	Pharmacologie
Mr Hervé **TILLY**	CB	Hématologie et transfusion
M. Gilles **TOURNEL**	HCN	Médecine Légale
Mr Olivier **TROST**	HCN	Chirurgie Maxillo-Faciale
Mr Jean-Jacques **TUECH**	HCN	Chirurgie digestive
Mr Jean-Pierre **VANNIER** *(surnombre)*	HCN	Pédiatrie génétique
Mr Benoît **VEBER**	HCN	Anesthésiologie - Réanimation chirurgicale
Mr Pierre **VERA**	CB	Biophysique et traitement de l'image
Mr Eric **VERIN**	HB	Service Santé Réadaptation
Mr Eric **VERSPYCK**	HCN	Gynécologie obstétrique
Mr Olivier **VITTECOQ**	HB	Rhumatologie
Mme Marie-Laure **WELTER**	HCN	Physiologie

MAITRES DE CONFERENCES DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mme Noëlle **BARBIER-FREBOURG**	HCN	Bactériologie – Virologie
Mme Carole **BRASSE LAGNEL**	HCN	Biochimie
Mme Valérie **BRIDOUX HUYBRECHTS**	HCN	Chirurgie Vasculaire
Mr Gérard **BUCHONNET**	HCN	Hématologie
Mme Mireille **CASTANET**	HCN	Pédiatrie
Mme Nathalie **CHASTAN**	HCN	Neurophysiologie
Mme Sophie **CLAEYSSENS**	HCN	Biochimie et biologie moléculaire
Mr Moïse **COEFFIER**	HCN	Nutrition
Mr Serge **JACQUOT**	UFR	Immunologie
Mr Joël **LADNER**	HCN	Epidémiologie, économie de la santé
Mr Jean-Baptiste **LATOUCHE**	UFR	Biologie cellulaire
Mr Thomas **MOUREZ** *(détachement)*	HCN	Virologie
Mr Gaël **NICOLAS**	HCN	Génétique
Mme Muriel **QUILLARD**	HCN	Biochimie et biologie moléculaire
Mme Laëtitia **ROLLIN**	HCN	Médecine du Travail
Mr Mathieu **SALAUN**	HCN	Pneumologie
Mme Pascale **SAUGIER-VEBER**	HCN	Génétique
Mme Anne-Claire **TOBENAS-DUJARDIN**	HCN	Anatomie
Mr David **WALLON**	HCN	Neurologie
Mr Julien **WILS**	HCN	Pharmacologie

PROFESSEUR AGREGE OU CERTIFIE

Mr Thierry **WABLE**	UFR	Communication
Mme Mélanie **AUVRAY-HAMEL**	UFR	Anglais

II - PHARMACIE

PROFESSEURS

Mr Thierry **BESSON**	Chimie Thérapeutique
Mr Roland **CAPRON** (PU-PH)	Biophysique
Mr Jean **COSTENTIN** (Professeur émérite)	Pharmacologie
Mme Isabelle **DUBUS**	Biochimie
Mr François **ESTOUR**	Chimie Organique
Mr Loïc **FAVENNEC** (PU-PH)	Parasitologie
Mr Jean Pierre **GOULLE** (Professeur émérite)	Toxicologie
Mr Michel **GUERBET**	Toxicologie
Mme Isabelle **LEROUX - NICOLLET**	Physiologie
Mme Christelle **MONTEIL**	Toxicologie
Mme Martine **PESTEL-CARON** (PU-PH)	Microbiologie
Mr Rémi **VARIN** (PU-PH)	Pharmacie clinique
Mr Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

MAITRES DE CONFERENCES

Mme Cécile **BARBOT**	Chimie Générale et Minérale
Mr Jérémy **BELLIEN** (MCU-PH)	Pharmacologie
Mr Frédéric **BOUNOURE**	Pharmacie Galénique
Mr Abdeslam **CHAGRAOUI**	Physiologie
Mme Camille **CHARBONNIER (LE CLEZIO)**	Statistiques
Mme Elizabeth **CHOSSON**	Botanique
Mme Marie Catherine **CONCE-CHEMTOB**	Législation pharmaceutique et économie de la santé
Mme Cécile **CORBIERE**	Biochimie
Mr Eric **DITTMAR**	Biophysique
Mme Nathalie **DOURMAP**	Pharmacologie
Mme Isabelle **DUBUC**	Pharmacologie
Mme Dominique **DUTERTE- BOUCHER**	Pharmacologie
Mr Abdelhakim **ELOMRI**	Pharmacognosie
Mr Gilles **GARGALA** (MCU-PH)	Parasitologie

Mme Nejla EL **GHARBI-HAMZA**	Chimie analytique
Mme Marie-Laure **GROULT**	Botanique
Mr Hervé **HUE**	Biophysique et mathématiques
Mme Laetitia **LE GOFF**	Parasitologie – Immunologie
Mme Hong **LU**	Biologie
M. Jérémie **MARTINET** (MCU-PH)	Immunologie
Mme Marine **MALLETER**	Toxicologie
Mme Sabine **MENAGER**	Chimie organique
Mme Tiphaine **ROGEZ-FLORENT**	Chimie analytique
Mr Mohamed **SKIBA**	Pharmacie galénique
Mme Malika **SKIBA**	Pharmacie galénique
Mme Christine **THARASSE**	Chimie thérapeutique
Mr Frédéric **ZIEGLER**	Biochimie

PROFESSEURS ASSOCIES

Mme Cécile **GUERARD-DETUNCQ**	Pharmacie officinale
Mr Jean-François **HOUIVET**	Pharmacie officinale

PROFESSEUR CERTIFIE

Mme Mathilde **GUERIN**	Anglais

ASSISTANT HOSPITALO-UNIVERSITAIRE

Mme Anaïs **SOARES**	Bactériologie

ATTACHES TEMPORAIRES D'ENSEIGNEMENT ET DE RECHERCHE

Mme Sophie **MOHAMED**	Chimie Organique

LISTE DES RESPONSABLES DES DISCIPLINES PHARMACEUTIQUES

Mme Cécile **BARBOT**	Chimie Générale et minérale
Mr Thierry **BESSON**	Chimie thérapeutique
Mr Roland **CAPRON**	Biophysique
Mme Marie-Catherine **CONCE-CHEMTOB**	Législation et économie de la santé
Mme Elisabeth **CHOSSON**	Botanique
Mme Isabelle **DUBUS**	Biochimie
Mr Abdelhakim **ELOMRI**	Pharmacognosie
Mr Loïc **FAVENNEC**	Parasitologie
Mr Michel **GUERBET**	Toxicologie
Mr François **ESTOUR**	Chimie organique
Mme Isabelle **LEROUX-NICOLLET**	Physiologie
Mme Martine **PESTEL-CARON**	Microbiologie
Mr Mohamed **SKIBA**	Pharmacie galénique
Mr Rémi **VARIN**	Pharmacie clinique
M. Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

III – MEDECINE GENERALE

PROFESSEUR DES UNIVERSITES MEDECIN GENERALISTE

Mr Jean-Loup **HERMIL** (PU-MG)	UFR	Médecine générale

MAITRE DE CONFERENCE DES UNIVERSITES MEDECIN GENERALISTE

Mr Matthieu **SCHUERS** (MCU-MG)	UFR	Médecine générale

PROFESSEURS ASSOCIES A MI-TEMPS – MEDECINS GENERALISTE

Mr Emmanuel **LEFEBVRE**	UFR	Médecine Générale
Mme Elisabeth **MAUVIARD**	UFR	Médecine générale
Mr Philippe **NGUYEN THANH**	UFR	Médecine générale
Mme Yveline **SEVRIN**	UFR	Médecine générale
Mme Marie Thérèse **THUEUX**	UFR	Médecine générale

MAITRE DE CONFERENCES ASSOCIE A MI-TEMPS – MEDECINS GENERALISTES

Mme Laëtitia **BOURDON**	UFR	Médecine Générale
Mr Pascal **BOULET**	UFR	Médecine générale
Mr Emmanuel **HAZARD**	UFR	Médecine Générale
Mme Lucile **PELLERIN**	UFR	Médecine générale

ENSEIGNANTS MONO-APPARTENANTS

PROFESSEURS

Mr Serguei **FETISSOV** (med)	Physiologie (ADEN)
Mr Paul **MULDER** (phar)	Sciences du Médicament
Mme Su **RUAN** (med)	Génie Informatique

MAITRES DE CONFERENCES

Mr Sahil **ADRIOUCH** (med)	Biochimie et biologie moléculaire (Unité Inserm 905)
Mme Gaëlle **BOUGEARD-DENOYELLE** (med)	Biochimie et biologie moléculaire (UMR 1079**)**
Mme Carine **CLEREN** (med)	Neurosciences (Néovasc)
M. Sylvain **FRAINEAU** (med)	Physiologie (Inserm U 1096)
Mme Pascaline **GAILDRAT** (med)	Génétique moléculaire humaine (UMR 1079)
Mr Nicolas **GUEROUT** (med)	Chirurgie Expérimentale
Mme Rachel **LETELLIER** (med)	Physiologie
Mme Christine **RONDANINO** (med)	Physiologie de la reproduction
Mr Antoine **OUVRARD-PASCAUD** (med)	Physiologie (Unité Inserm 1076)
Mr Frédéric **PASQUET**	Sciences du langage, orthophonie
Mr Youssan Var **TAN**	Immunologie
Mme Isabelle **TOURNIER** (med)	Biochimie (UMR 1079)

CHEF DES SERVICES ADMINISTRATIFS : Mme Véronique **DELAFONTAINE**

HCN - Hôpital Charles Nicolle	*HB - Hôpital de BOIS GUILLAUME*
CB - Centre Henri Becquerel	*CHS - Centre Hospitalier Spécialisé du Rouvray*
CRMPR - Centre Régional de Médecine Physique et de Réadaptation	*SJ – Saint Julien Rouen*

Identification de maladies associées aux régressions du développement dans le syndrome de Phelan-McDermid: analyse exhaustive d'un registre international

Le domaine de l'informatique médicale s'attache à permettre une utilisation optimale des données de santé, tant pour la prise en charge médicale des patients que pour la recherche médicale. L'informatisation des données de santé permet une facilitation de l'accès à ces données et offre l'opportunité de les exploiter de façon automatisée.

L'acquisition de données est un facteur limitant important de la recherche clinique. Pour de nombreux travaux de recherche, l'objectif et la méthodologie sont clairement établis, mais la création de cohortes de patients correspondant aux critères de l'étude est difficile, et parfois rendue impossible par le manque de moyens et la complexité du recrutement, particulièrement lorsque la pathologie étudiée est rare.

Recruter des patients pour réaliser des études cliniques portant sur des maladies orphelines est extrêmement complexe, limitant souvent les chercheurs à des cohortes de quelques patients. Les registres de maladies peuvent pallier ce problème de recrutement, en collectant des données de patients atteints d'une pathologie en amont du travail de recherche. Contrairement aux données des services hospitaliers, pour lesquels les données sont destinées uniquement au soin des patients qui les fréquentent, les registres peuvent collecter des informations de façon plus complète sur une pathologie en se focalisant sur des questions susceptibles d'être utiles dans un cadre de recherche clinique, et centraliser ces données de façon internationale sans avoir à affronter les nombreux problèmes d'interopérabilité des systèmes informatiques hospitaliers.

Grâce à ces registres il est possible de créer des entrepôts de données destinées à la recherche, et de permettre aux chercheurs de recruter virtuellement un nombre de patients très supérieur à ce qui aurait été réalisable dans une étude clinique standard.

Le syndrome de Phelan-McDermid est une de ces maladies orphelines pour lesquelles la recherche clinique est très difficile à mettre en oeuvre. Cette maladie génétique consiste en une altération du gène *SHANk3*, situé sur l'extrémité du bras court du chromosome 22. On estime qu'elle concerne environ 2000 patients dans le monde.
Le type d'altération génétique causant le syndrome de Phelan-McDermid est très variable, il peut s'agir d'une mutation, une micro délétion ne touchant que le gène *SHANK3*, une délétion interstitielle de taille variable ou une large délétion terminale de l'extrémité 22q13, touchant jusqu'à 140 gènes. Son expression clinique est donc elle aussi extrêmement variée. Parmi les symptômes les plus fréquemment retrouvés, on retrouve l'hypotonie, le retard du développement psychomoteur, un retard intellectuel variable, un syndrome dysmorphique, des troubles du spectre autistique, l'épilepsie. Les manifestations commencent généralement rapidement dans l'enfance mais il existe des cas diagnostiqués à l'âge adulte.

Les régressions du développement, c'est à dire la perte d'un développement psychomoteur auparavant acquis, sont une complication fréquente et particulièrement sévère du syndrome de Phelan-McDermid [1]. Elles concernent entre 30 et 75 % des patients selon les études, et peuvent toucher la motricité fine et globale [1, 2, 3], le langage [4], les compétences sociales. Ces pertes de compétences acquises peuvent:

- entrainer une réduction d'autonomie importante avec un retour à une dépendance pour les actes de la vie quotidienne
- réduire de façon significative les possibilités de communication
- entrainer des complications médicales potentiellement sévères notamment par l'apparition

de troubles de la déglutition pouvant mener à des pneumopathies d'inhalation ou une asphyxie causée par une inhalation [2].

Le mécanisme de ces régressions développementales est inconnu [5], et l'on ne dispose pas de moyen de prédire, prévenir ou traiter leur apparition. La définition même des régression du développement n'est pas standardisée [6].

Plusieurs travaux ont étudié ou décrit les manifestations des regressions chez les patients atteints du syndrome de Phelan-McDermid:
Lors de réunions d'un groupe de soutien de patients atteints du syndrome de Phelan-McDermid, 17/48 parents (35%) on rapporté la survenue de régression du développement de leur enfant. Cependant, le type de régression et l'âge de survenue n'étaient pas précisés [5]. Une étude réalisée en 2010 et incluant 13 patients rapportait deux sujets (15%) ayant présenté une régression développementale, sans spécifier la méthode d'évaluation ou le type de compétence affectée [7]. Dans une autre étude publiée en 2013 sur 32 patients, neuf (28%) présentaient une perte du développement. Des question du Autism Diagnostic Interview-Revised (ADI-R), un questionnaire destiné aux patients atteints d'un trouble autistique et contenant des questions évaluant les régressions du développement [8]. Les régressions concernaient principalement le langage, et l'âge de survenue variait de 15 mois à 17 ans. Une étude sur sept patients retrouva une perte progressive du langage, des compétences motrices et sociales pour tous les patients. Cependant, la méthode d'évaluation n'était pas spécifiée [9].
Une autre étude retrouvait une régression pour 18 / 42 patients (43%) [1] . Dans une étude portant sur 11 patients, 4 (36%) avaient présenté une régression développementale touchant le langage pour deux patients, et les compétences motrices pour deux patients [10].
Dans un rapport de cas publié en 2013, un patient avait présenté une perte du langage à l'âge de 15 mois. Une régression des compétences motrices et des aptitudes d'auto assistance étaient survenus à l'âge de deux ans et demi [11]. Dans une autre étude de cas publiée en

2015, deux patients avaient présenté une régression aux âges de 12 et 13 ans respectivement. Tous les deux avaient une perte de compétences motrices et une perte d'autonomie, et avaient présenté un syndrome catatonique, et l'un des patients avait présenté une perte du langage [12]. Dans un rapport de cas incluant six patients atteints de crises d'épilepsie associées au syndrome de Phelan-McDermid [4], un patient avait présenté une régression du langage à l'âge de quatre ans. Ces troubles ne s'étaient jamais améliorés, et des troubles sphinctériens étaient apparus à l'âge de 16 ans. Un autre patient rapporté dans cette étude avait présenté des troubles sphinctériens à l'âge de 10 ans. L'étude de l'épilepsie de ces patients n'avait pas permis de retrouver d'association avec les régressions du développement . Dans un autre rapport de cas [13], les crises d'épilepsie étaient survenues avant les régressions développementales chez 2 / 3 patients. Le premier patient avait perdu des acquisistion motrices et du langage à environ six ans dans un contexte d'augmentation de l'activité épileptique. Les capatités motrices avaient été récupérées après obtention d'un meilleur controle de l'épilepsie. Le second patient avait présenté un premier épisode de régression du langage à l'âge de deux ans, et à l'âge de sept ans une maladie épileptique avait été diagnostiquée et avait perdu des capacités motrices dans la même période. Un rapport [14] de huit patients publié en 2008 montrait une perte de langage à l'âge de 15 mois pour deux sujets, avec une récupération après plusieurs mois, et une perte de compétences sociales était reportée durant l'adolescence pour trois patients. Une étude clinique [15] portant sur les dysfonctions mitochondriales des patients atteints du syndrome de Phelan-McDermid retrouvait 22/30 (74%) patients ayant présenté une régression du développement. Les régressions concernaient le plus souvent le langage (16/22, 73%), et les pertes de capacités motrices et sociales étaient moins fréquentes (10/22, 45% and 7/22, 32% respectivement). Ils ont montré que la présence d'une régression faisait partie des signes associés à la présence d'une activité anormale des complexes de la chaîne de transport d'électrons. Une étude publiée en 2018 [16] rapportait 11/17 (65%) patients ayant présenté une régression du développement. Un épisode aigu était noté chez trois patients (deux processus infectieux et un épisode épileptique). La perte de langage était le symptome de régression le plus fréquemment rapporté.

La revue des travaux de recherche existants montre que les manifestations cliniques sont extrêmement variées, les régression du développement peuvent apparaître dans différent domaines des acquisitions et leur description est souvent imprécise. Comme la maladie est extrêmement rare, le nombre de patients inclus est en général faible et souvent inférieur à 20, ce qui rend l'étude de ces manifestations particulièrement difficile.
La Phelan-McDermid Syndrome Fundation (PMSF) est une association fondée par des familles de patients atteints du syndrome de Phelan-McDermid. Elle tient notamment le Phelan-McDermid International Registry (PMSIR), un registre international de patients atteints du PMS. Ce registre comprend des informations issues de trois sources de données:

- Un questionnaire remplissable directement en ligne par les familles des patients, traitant de questions relatives au diagnostic, à la présence d'autres pathologies, les traitements et interventions réalisées, le développement psychomoteur.
- Les compte rendus des analyses génétiques ayant servi au diagnostic, incluant Comparative Genome Hybridization arrays, Single Nucleotide Polymorphism array, et microarray, traitées par des conseillers en génétique avant ajout dans le registre.
- Les courriers médicaux numérisés, recueillis par un opérateur indépendant auprès des différents établissements hospitaliers fréquentés par les patients, après obtention du consentement des familles.

Afin de pouvoir être exploitées par des chercheurs, les données de ce registre ont été incluses dans une base de données, le Phelan-McDermid Syndrome Data Network (PMS_DN).
Cet entrepôt au format i2b2 permet d'intégrer aisément des données d'origine multiples, grâce à la structure de la base de données [17], le "star-schema". Ce modèle, créé spécifiquement pour le stockage de données de patients, se base sur le concept "entité-attribut-valeur" pour permettre une flexibilité dans l'import des données ainsi que dans le types des données importées. Cela

permet d'alléger le travail d'extraction et de transformation avant le chargement dans la base, ainsi que le chargement de données d'origine fondamentalement différente sans modifier la structure de la base de données.

PMS_DN contient les réponses aux questionnaires, les données des compte rendus génétiques sous la forme de 57 champs structurés, et les compte rendus cliniques traités par reconnaissance optique de caractères, dé-identification et extraction de concepts par traitement automatique de la langue (TAL). Ce processus de reconnaissance optique de caractères suivi de TAL a permis d'intégrer les compte rendus cliniques sous la forme de Concept Unique Identifier (CUI) du métathesaurus de l'Unified Medical Language System (UMLS). Les CUIs ont ensuite été traduits dans 20 terminologies médicales en utilisant les correspondances établies par l'UMLS, permettant l'expression des concepts dans des terminologies et ontologies adaptées à différent contextes. Cette annotation des compte rendus clinique permet d'exploiter ces dernières avec des outils statistiques classiques, sans avoir besoin de réinterpréter leur contenu en texte libre.

Objectif

L'objectif de ce travail était d'utiliser les informations disponibles dans cette base de données afin d'identifier des pathologies liées à l'apparition de régressions du développement chez les patients atteints du syndrome de Phelan-McDermid.

References

[1] G. Reierson, J. Bernstein, W. Froehlich-Santino, A. Urban, C. Purmann, S. Berquist, J. Jordan, R. O'Hara, and J. Hallmayer, "Characterizing regression in phelan mcdermid syndrome (22q13 deletion syndrome).," *J Psychiatr Res*, Aug 2017.

[2] V. Hughes, "Scientists track adult regression in autism-related syndrome." Jul 2012.

[3] M. H. Willemsen, J. H. M. Rensen, H. M. J. van Schrojenstein-Lantman de Valk, B. C. J. Hamel, and T. Kleefstra, "Adult phenotypes in angelman- and rett-like syndromes.," *Mol*

Syndromol, Apr 2012.

[4] M. G. Figura, A. Coppola, M. Bottitta, G. Calabrese, L. Grillo, D. Luciano, L. Del Gaudio, C. Torniero, S. Striano, and M. Elia, "Seizures and eeg pattern in the 22q13.3 deletion syndrome: clinical report of six italian cases.," *Seizure*, Oct 2014.

[5] S. Wilson, A. Djukic, S. Shinnar, C. Dharmani, and I. Rapin, "Clinical characteristics of language regression in children.," *Dev Med Child Neurol*, Aug 2003.

[6] D. Zhang, F. Bedogni, S. Boterberg, C. Camfield, P. Camfield, T. Charman, L. Curfs, C. Einspieler, G. Esposito, B. De Filippis, R. P. Goin-Kochel, G. U. Hoglinger, D. Holzinger, A.-M. Iosif, G. E. Lancioni, N. Landsberger, G. Laviola, E. M. Marco, M. Muller, J. L. Neul, K. Nielsen-Saines, A. Nordahl-Hansen, M. F. O'Reilly, S. Ozonoff, L. Poustka, H. Roeyers, M. Rankovic, J. Sigafoos, K. Tammimies, G. S. Townend, L. Zwaigenbaum, M. Zweckstetter, S. Bolte, and P. B. Marschik, "Towards a consensus on developmental regression.," *Neurosci Biobehav Rev*, Aug 2019.

[7] S. U. Dhar, D. del Gaudio, J. R. German, S. U. Peters, Z. Ou, P. I. Bader, J. S. Berg, M. Blazo, C. W. Brown, B. H. Graham, T. A. Grebe, S. Lalani, M. Irons, S. Sparagana, M. Williams, J. A. r. Phillips, A. L. Beaudet, P. Stankiewicz, A. Patel, S. W. Cheung, and T. Sahoo, "22q13.3 deletion syndrome: clinical and molecular analysis using array cgh.," *Am J Med Genet A*, Mar 2010.

[8] L. Soorya, A. Kolevzon, J. Zweifach, T. Lim, Y. Dobry, L. Schwartz, Y. Frank, A. T. Wang, G. Cai, E. Parkhomenko, D. Halpern, D. Grodberg, B. Angarita, J. P. Willner, A. Yang, R. Canitano, W. Chaplin, C. Betancur, and J. D. Buxbaum, "Prospective investigation of autism and genotype-phenotype correlations in 22q13 deletion syndrome and shank3 deficiency.," *Mol Autism*, Jun 2013.

[9] A. Denayer, H. Van Esch, T. de Ravel, J.-P. Frijns, G. Van Buggenhout, A. Vogels, K. Devriendt, J. Geutjens, P. Thiry, and A. Swillen, "Neuropsychopathology in 7 patients with the 22q13

deletion syndrome: Presence of bipolar disorder and progressive loss of skills.," *Mol Syndromol*, Jun 2012.

[10] M. A. Manning, S. B. Cassidy, C. Clericuzio, A. M. Cherry, S. Schwartz, L. Hudgins, G. M. Enns, and H. E. Hoyme, "Terminal 22q deletion syndrome: a newly recognized cause of speech and language disability in the autism spectrum.," *Pediatrics*, Aug 2004.

[11] M. Macedoni-Luksic, D. Krgovic, B. Zagradisnik, and N. Kokalj-Vokac, "Deletion of the last exon of shank3 gene produces the full phelan-mcdermid phenotype: a case report.," *Gene*, Jul 2013.

[12] S. Serret, S. Thümmler, E. Dor, S. Vesperini, A. Santos, and F. Askenazy, "Lithium as a rescue therapy for regression and catatonia features in two SHANK3 patients with autism spectrum disorder: case reports," *BMC Psychiatry*, vol. 15, p. 107, May 2015.

[13] D. M. Cochoy, A. Kolevzon, Y. Kajiwara, M. Schoen, M. Pascual-Lucas, S. Lurie, J. D. Buxbaum, T. M. Boeckers, and M. J. Schmeisser, "Phenotypic and functional analysis of shank3 stop mutations identified in individuals with asd and/or id.," *Mol Autism*, 2015.

[14] A. Philippe, N. Boddaert, L. Vaivre-Douret, L. Robel, L. Danon-Boileau, V. Malan, M.-C. de Blois, D. Heron, L. Colleaux, B. Golse, M. Zilbovicius, and A. Munnich, "Neurobehavioral profile and brain imaging study of the 22q13.3 deletion syndrome in childhood.," *Pediatrics*, Aug 2008.

[15] R. E. Frye, D. Cox, J. Slattery, M. Tippett, S. Kahler, D. Granpeesheh, S. Damle, A. Legido, and M. J. Goldenthal, "Mitochondrial dysfunction may explain symptom variation in phelan-mcdermid syndrome.," *Sci Rep*, Jan 2016.

[16] S. De Rubeis, P. M. Siper, A. Durkin, J. Weissman, F. Muratet, D. Halpern, M. D. P. Trelles, Y. Frank, R. Lozano, A. T. Wang, J. L. J. Holder, C. Betancur, J. D. Buxbaum, and A. Kolevzon, "Delineation of the genetic and clinical spectrum of phelan-mcdermid syndrome caused by shank3 point mutations.," *Mol Autism*, 2018.

[17] J. G. Klann, A. Abend, V. A. Raghavan, K. D. Mandl, and S. N. Murphy, “Data interchange using i2b2.,” *J Am Med Inform Assoc*, Sep 2016.

Identifying medical conditions associated with developmental regressions in the Phelan-McDermid syndrome: extensive analysis of an international registry

Contents

1 Introduction

1.1 Phelan-McDermid Syndrome

The Phelan-McDermid Syndrome (PMS) or *q22q13* deletion syndrome is a rare neuropsychiatric disorder, with approximately 2,000 cases diagnosed worldwide, caused by a deletion of the terminal end of the chromosome 22 or a point mutation in *SHANK3* gene [1].

Regression, the loss of acquired skills, is a frequent symptom reported by PMS patients [2]. The clinical evolution of these regressions is marked by an acute loss of certain functions or the deterioration of a previously controlled condition. They are a frequent issue of several neurodevelopmental disorders, such as autism spectrum disorders and RETT syndrome [3, 4, 5, 6] but little is known about their underlying mechanism [7].

Developmental regressions in PMS can affect motor/cognitive skills [2, 8, 9], language skills [10] and self-help skills, that can lead to life-threatening complications, such as aspiration due to the occurrence of swallowing difficulties [8]. They have a major impact on the quality of life on patients with PMS, and understanding their underlying mechanism could help to prevent them or reduce their impact on the life of the patients.

The research efforts focusing on PMS are severely impaired by the lack of patient data. Recruiting patients to perform medical studies is difficult, and most of the time only a few patients can be included by the investigators. Therefore, an important part of the available knowledge about regressions in PMS is represented by case reports or studies including a low number of patients. This limits the ability to discover patterns and to perform inferential analyses due to a low statistical power.

There is no standard definition of developmental regression [11], and the type of affected skills varies widely across individual patients. Researchers studying regression in PMS use their own definitions of regression, sometimes using questions related to regression in the the Autism Diagnostic Interview-Revised (ADI-R), a diagnostic instrument for assessing autism [12].

Due to this lack of standardisation, meta-analyses focused on regressions in PMS patients are difficult to conduct, and the estimated prevalence of regression varies between studies.

During a support group for parents with 22q13 deletion, 17/48 parents (35%) reported regression happening to their children; however, the types of regression and age of onset were not defined [7]. A study with 13 patients reported two individuals (15%) with cases of regression, without specifying the method of evaluation or the type of affected skills [13]. In another study published in 2013 on 32 patients, nine patients (28%) reported loss of skills. Questions from the ADI-R were used to assess regression [14]. The regressions mostly affected the language skills, and age of onset varied from 15 months to 17 years. One study on seven patients retrieved a progressive loss of language, gross and fine motor skills and social skills for all of the subjects. However, the method of evaluation was not specified [15].

An additional study reported 18 out of 42 patients (43%) to have regression present [2], and in another study on 11 patients, 4 (36%) presented with developmental regression (language skills loss for two patients, motor skills for two patients) [16].

In a case report published in 2013, a patient presented language skills loss at 15 months of age. Regression of motor and self-help skills were present at 2.5 years of age [17]. In another case report published in 2015, two patients presented regression at age 12 and 13. Both had motor and self-help skills affected and presented catatonia symptoms, and one of the patients presented language regression [18]. In a case report including six patients presenting with EEG abnormalities and PMS and epilepsy in three of the patients, one was described as having a loss of language at the age of four. These language skills were never recovered, and a loss of sphincter control appeared at age 16. Another patient in this study presented a loss of sphincter control at age ten. The other patients did not present a developmental regression. The study of epilepsy on the six patients did not retrieve an association between seizure and regression [10]. In another case report [19], seizures were found to occur before a regression in two out of three patients. The first patient lost language and motor skills at around 6 years of age in

a context of increased seizure activity. Motor skills were regained when epilepsy was better controlled. The second patient had a first episode of language regression at two years of age, and at the age of seven, he was diagnosed with epilepsy, and lost motor skills at the same period of time. A 2008 report of eight patients showed a loss of language skills after 15 months for two subjects, with a recovery after several months, and a loss of social skills was recorded during the teenage years for three patients [20]. A clinical study focusing on mitochondrial dysfunction in patients with PMS reported 22/30 (74%) patients with a developmental regression. Regression of language skills was the most common (16/22, 73%), and loss of social and motor skills were less common (10/22, 45% and 7/22, 32% respectively). They showed that having symptoms of ASD, developmental regression, failure to thrive and/or exercise intolerance/fatigue was associated with a greater likelihood of abnormal ETC complex activity. [21]. A study published in 2018 reported 11/17 (65%) patients with developmental regression. An acute event was noted in three patients (2 infectious diseases and 1 seizure). Language loss was the most commonly reported regression [22].

There are no identified factors determining the risk of presenting with a regression during the natural evolution of PMS. Some comorbidities are suspected to be a possible trigger for the regressions. A study on seven patients retrieved loss of skills after acute events such as septic shock, epileptic state, catatonic phase, or malignant neuroleptic syndrome [15]. Seizures are considered to be a possible risk factor of regression in other neurodevelopmental diseases. However, a recent study on 50 patients with PMS was not able to show an association between seizures and the risk of regression [2].

The association between seizures and developmental regression is highly discussed in patients with autism spectrum disorders, but the nature of the relationship, if any, has not been established yet, and the understanding of the mechanisms of developmental regression remain very poor [23]. Deonna et al. [24] published a case report of two children with autism where a temporal relation between the seizures and developmental regressions. However, this temporal

association is not always present and the relationship between autism, developmental regression, and epilepsy is still unclear [25]. Viscidi et al. [26], in a study including 5,815 patients with autism, found an association between autism and developmental regression ($p<0.001$), becoming non significant ($p=0.6$) when adjusting on variables including intelligence quotient and age.

The Rett syndrome is a neurodevelopmental disease where developmental regression always occur, and epileptic seizures are present among 48 to 94% of the patients [27].

There are no treatments for developmental regressions occuring in PMS. A study evaluated the effect of lithium to reverse the clinical symptoms with success on two patients [18]. Further studies are needed to confirm the effect of this medication on regressions.

Increasing the knowledge about the causes and risk factors of regression in PMS could help the development of preventive / curative therapeutics and increase the quality of life of these patients.

1.2 Phelan-McDermid Syndrome International Registry

The PMS Foundation (PMSF, 22q13.org) is a nonprofit foundation founded and run by PMS families that promotes awareness and research of PMS. Patient data from hundreds of families of PMS patients were collected worldwide, with their informed consent and stored in a registry, the Phelan-McDermid Syndrome International Registry (PMSIR, pmsiregistry.patientcrossroads.org).

Research on rare disorders is hampered by the paucity of available data, a consequence of the small patient population size. In order to facilitate research into rare diseases, it is crucial to devise ways to enable maximal access to patient's data.

The PMSIR fills this role by collecting the as much data as possible for each patient, from various sources: Parent Reported Outcomes (PROs), filled by the patient's primary caregivers, clinical notes from all the medical institutions visited by the patients, and curated genetic reports.

1.2.1 Parent Reported Outcomes

Three questionnaires (Clinical, Development, and Adult) filled by parents and/or caregivers of PMS patients were included in the database. These forms contain answers to approximately 1,300 questions about clinical (diagnoses, procedures, lab tests, medications, patient behavior, conditions) and developmental features. Some questions ask about manifestations that are specific to PMS, and others ask about common symptoms, and conditions that occur in Autism Spectrum Disorder.

1.2.2 Genetic Reports

Reports from genetic tests, including Comparative Genome Hybridization (CGH) arrays, Single Nucleotide Polymorphism (SNP) arrays and microarrays are also collected.
These reports were manually curated by genetic counselors, who filled 57 structured fields to represent genetic abnormalities, before being added to the PMSIR.

1.2.3 Clinical Notes

The families of patients gave their informed consent to a third-party vendor, CareSync (caresync.com), to request their health records on their behalf from healthcare providers.
The entirety of the clinical notes were collected from each hospital visited by the patients, and stored as PDFs documents in the registry.

1.3 Phelan-McDermid Syndrome Data Network

The Phelan-McDermid Syndrome Data Network (PMS_DN) is a Patient Powered Research Network [28] funded by the Patient Centered Outcomes Research Institute (PCORI, www.pcori.org). The patients and their families are the primary stakeholders, and manage all aspects of data governance. This initiative furthers Precision Medicine research into PMS by providing access to the PROs, clinical notes and genetic data provided by the PMSF from one single, standardised

database [29]. All the data, at the patient level, are hosted by a HIPAA compliant [30] cloud provider (Amazon Web Services). Authorized researchers can securely connect to the database, allowing for several studies to be performed on the same dataset, and eliminating the need for data curation for every new research work.

The data issue from the Clinical Notes were stored as PDFs in the PMSIR. In order to represent the knowledge available in a way that allows statistical analysis, these PDFs have been processed by an Optical Character Recognition (OCR) tool in order to extract the raw text content, and a Natural Language Processing (NLP) software, cTakes [31] identified occurrences of Unified Medical Language System (UMLS) [32] concepts from the clinical notes. The UMLS concepts were mapped to 20 medical terminologies including ICD-9, SNOMED CT, Human Phenotype Ontology, utilizing the diversity and expressiveness of each of these terminologies to represent the different medical concepts identified in these clinical notes.

1.4 Objective

The goal of this study was to identify medical conditions associated with developmental regression in PMS, using PMS_DN, the largest database of PMS patients, to understand further the mechanisms of developmental regressions in the Phelan-McDermid syndrome.

2 Materials & Methods

2.1 Patients

All the patients participating to the PMSIR were condidered for inclusion.
Patients for which genetic test had not been properly validated in the registry were excluded from the study to ensure that all the patients had been diagnosed with mutation or deletion affecting SHANK3. Patients for which no data about the presence of a developmental regression in either the questionaires or the clinical not were not included in the analysis.

2.2 Data preparation and feature selection

2.2.1 PRO

The Parent Reported Outcomes provided an important number of answers on various medical conditions. However, this type of information relies on the parents/patients being able to fill a maximum of questions. The quality of these datasets is therefore often impaired by a significant amount of missing values [33].
the PMS Data Network's questionnaires contains a very large amount of questions, making it challenging for the parent to fill completely. In this setting, data imputation becomes infeasible and it is necessary to choose a statistical analysis capable to handle a vast amount of missing values. In order to keep only the questions relevant to our research, we reviewed the questions and selected the ones that assessed the presence or absence of a medical condition, filtering out questions in other topics as well as questions related to the presence of a developmental regression.

2.2.2 Clinical Notes

The clinical notes do not contain missing value by definition, the absence of a term being used as the absence of the condition. This allows us to run more refined statistical analyses. To

represent the clinical concepts, we chose to use the SNOMED CT terminology. Among all the terminologies availables in the database, the SNOMED CT has the advantage of being a clinically oriented and generalist terminology. As opposed to the ICD-9 and ICD-10 terminologies, which were designed with medical billing in mind, it's goal is to be used by Electronical Health Records to accurately represent the health conditions of each patient. The SNOMED CT is also extremely precise as it counts more than 300,000 codes and is able to represent medical conditions in a very accurate way. It is the terminology that contains the highest number of concepts in the database, with more than 13,000 different codes.

Since we are focusing on medical conditions, we selected only the terms in the "Clinical Findings" branch of the SNOMED CT.

Since we have much more clinical concepts than patients, the analysis of this dataset was challenging. In such a situation, the standard statistical analysis methods cannot be used reliably anymore [34]. Therefore, finding a way to meaningfully reduce the amount of features is an important factor to increase the relevance of the results, as well as using adapted statistical analyses.

One efficient way to reduce the dimensionality of clinical terms is to express them in a more general way by reducing the granularity of their representation. Instead of considering different types of a disease as individual variables, we can group them in one single entity, thus reducing the number of variables while maintaining coherence and meaningfulness of the phenotypes. We used the SNOMED CT's tree hierarchy to perform this transformation by aggregating the different medical concepts based on the length of their path in the tree. The deepest a term sits in the terminology tree, the more accurate and specific it is. By capping the codes to a maximal depth, we can map the codes that are too precise to their parent code. We then merge the codes by marking the presence of a diagnosis if any of the children codes had been diagnosed. We thus end up with a dataset containing fewer, shallower codes.

The stronger the dimensionality reduction, the broader and more general the information of the dataset. Thus deciding the ideal path length for this aggregation is not an easy task as it

requires choosing the best balance between accuracy of the data and it's compactness. Since the goal of the analysis is exploratory, we decided to run our analysis with all possible path lengths to not discard potentially interesting results.

2.3 Evaluation of the OCR and NLP processing

The processing step of the clinical notes by OCR and NLP was evaluated to ensure the medical codes resulting from this automatic processing of the clinical notes was accurate enough to proceed with the statistical analysis. We read the clinical history from each patient and manually annotated the presence of eleven medical conditions, and compared it to the concepts extracted by cTakes.

The concepts used for this evaluation were selected based on a variety of expected characteristics: some were expected to have a relatively high prevalence (hypotonia, gastroesophageal reflux disease), some were expected to be easy to detect by the NLP algorithm because of the unambiguous jargon (cardiac murmur, strabism, flat feet, asthma, pica, ptosis, lymphedema), and some were expected to be more difficult to interpret (sleep disorders, developmental regression) because they can be expressed in complex ways by the medical doctors. Some of the terms were medical conditions that appear independently of the presence of PMS, and others were comorbidities that are frequently associated with PMS, thus covering a wide range of situations.

We then computed the recall, precision and F1-score to evaluate the performance of the natural language processing for each concept. We synthetized the performances by calculating the macro-averaged values of these metrics.

2.4 Determination of the presence of a developmental regression

Since we have access to two complementary sources of data, the PROs and the Clinical Notes, determining the presence of a developmental regression can be done by computing this knowledge into one single feature.

2.4.1 PRO

The information from the PRO allowed us to determine the presence of a developmental regression for a large part of the patients.
Five questions relative to the presence of a developmental regression can be found in the "developmental" section of the questionnaires:

- Has The Patient Ever Experienced Regression In Cognitive Abilities?
- Has The Patient Ever Experienced A Loss Of Or Regression In A Previously Acquired Fine Motor Skill?
- Has The Patient Ever Experienced A Loss Of Or Regression In A Previously Acquired Gross Motor Skill?
- Has The Patient Ever Experienced A Loss Of Or Regression In A Previously Acquired Self-Help Skill?
- Has The Patient Ever Experienced Regression In Social Abilities?

The possible answers were "Yes", "Yes (by imputation)", "No", "No (by imputation)", "Unsure", "Not applicable".
The answer was considered positive if the parent entered "Yes" or "Yes (by imputation)", negative otherwise.
A patient was considered as having presented with a developmental regression if the answer to at least one of the answers was positive.
When some of the questions were not answered, only the answered questions were taken into consideration.

2.4.2 Clinical Notes

The mappings of the clinical codes corresponding to a developmental regression have been used to determine if a patient had a developmental regression (SNOMED CT code CO609225004 :

"Developmental Regression", HPO code HP:0002376 "Developmental Regression", OMIM code 618088 : "loss of speech").
If no code related to a developmental regression was found in the clinical notes, the patient was considered as not having presented a regression.

2.4.3 Creation of the combined feature

the presence of developmental regression was determined by combining the knowledge in the PRO and the Clinical Notes.
Patients for whom information of PRO and clinical notes was available were considered as having a developmental regression if any of the source was positive.

2.5 Statistical Analysis

The goal of the statistical analysis was to develop insight of the potential influencial factors of the presence of developmental regression.
We analyzed the PRO and the clinical notes separately, and used the results of the analysis to identify potential conditions associated with developmental regression.

2.5.1 PRO

In order to facilitate the analysis and try to detect as many associations as possible despite the low overall answer rate in the Parent Reported Outcomes, we performed a variant of Phenome-Wide Association Study (PheWAS) .
PheWAS [35] is typically used to detect the association of a genetic alteration against different phenotypes. In our analysis, in lieu of a genetic alteration, the cases and controls are determined by a phenotype: the presence or absence of a developmental regression, as defined by our previously described combined feature.
The phenotypes we tested against are the answers to the different answers from the questionnaires. For each of these phenotypes, a logistic regression adjusted on sex and age was

performed. The p-values were corrected for multiple testing using the Bonferroni correction, and a threshold of 0.05 on the corrected p-values was used as significant threshold.
The main advantage of this approach lies in the realisation of independent tests for each predictor, thus dampening the impact of the low answer rate.

2.5.2 Clinical Notes

For the data exploration and analysis of the Clinical Notes, we explored several approaches: a PheWAS, an Elastic Net-regularized logistic regression, Random Forest modelisation and a Gradient Boosting Machine. We compared their results and strengths/weaknesses, and tried to isolate a common signal that could show an association between developmental regression and a medical condition.
The first technique used was a PheWAS, identical to the one used for the PROs. It has the advantage of being computationnaly inexpensive, and, although it performs multiple independent tests, it gives a good overview of the potential statistical associations. It remains statistically coherent even with a very high number of variables.
The second approach consisted in a logistic regression with Elastic net regularization. The elastic net is a combination of ridge (L2) and lasso (L1) regularization. It's goal in to simplify the model by reducing the value of the least contributing variables. With the lasso component, unimportant features' coefficients are set to zero, thus resulting in feature selection by excluding those features from the model. This allows us to integrate the whole dataset in a single model, and identify the key features associated to the presence of a developmental regression in PMS. A grid search on a set of hyper-parameters with a 5 folds cross-validation, repeated 10 times, was performed to select the best values of alpha (balance between ridge and lasso regularization) and lambda (strength of the regularization). During the training, the model's performance was evaluated using the area under the ROC (AUC). The features selected by the final model were retained as possible associations with developmental regression.
The third analysis was a Random Forest. This machine learning algorithm is robust to highly

dimensional data. A grid search on a set of hyper-parameters with a 5 folds cross-validation, repeated 10 times, was performed to select the best number of features selected to build each tree. The model comprised 1,000 trees. During the training, the model's performance was evaluated using the area under the ROC (AUC). Feature importance was used on the final model to retain the possible associations with developmental regression.

The last model we used was xgboost [36], a gradient boosted trees implementation. This algorithm uses a sequence of small decision trees, each new tree reducing the loss function of the previous. It is very resilient to overfitting and can thus handle highly dimensional data. A grid search on a set of hyper-parameters with a 5 folds cross-validation, repeated 10 times, was performed to select the best learning rate, number of iterations and maximum tree depth. During the training, the model's performance was evaluated using the area under the ROC (AUC). Feature importance was used on the final model to retain the possible associations with developmental regression.

2.6 Exploration of the results: review of the Clinical Notes and PRO

The statistical analyses run on the PROs and the Clinical notes were used as a way to mine the vast amount of features and explore their possible association with developmental regression. After reviewing and comparing their results, we explored the relevance of these results and tried to assert their veracity by reviewing the information present in the PMS Data Network.

2.6.1 Review of the Clinical Notes

The critical point of the integration of the clinical notes was their transformation into UMLS CUIs, requiring OCR and natural language processing. With the results of the analysis, we could perform a more systematic and thorough exploitation of these documents. We read the entirety of the documents and explored in detail, for each patient, the results found as most relevant from the analyses.

With the full text of the notes taken by the Medical Doctors over the different hospitalisations of

the patients, we were able to draw stronger conclusions. In order to give arguments regarding the causal relationship of the results, we recorded and analyzed the temporality between the apparition of developmental regression and the results.

2.6.2 PRO

In the PRO, we further study the answers given to questions related to the important features of the different models.

3 Results

3.1 Included Patients

From the 99 patients for which the medical notes were available, 19 patients were excluded from the analysis because the results of the genetic test were not available or invalidated the suspiscion of Phelan-McDermid syndrome. From the remaining 80 patients, 2 were excluded because no information about the presence of a developmental regression was available.
In total, 78 patients were included for the Clinical Notes analysis.
Out of the 414 patients who answered questions from the PROs, 181 patients were excluded from the analysis because the results of the genetic test were not available or invalidated the suspiscion of Phelan-McDermid syndrome. From the remaining 233 patients, 36 were excluded because no information about the presence of a developmental regression was available, either in the PROs or the Clinical Notes.
In total, 197 patients were included for the PRO analysis (**Table 1**).

3.2 Demographics

in the patients included in the PRO analysis, the average age was 11.3 (SD: 7.9, median: 8.9). The age in the clinical notes averaged to 11.7 (SD: 6.4, median: 10.3).
The sex ratio was 0.8 (90/107) in the PRO population, 1.2 (43/35) in the Clinical Notes population.

Table 1: Demographics

		Developmental Regression	
		No	Yes
n		110	196
Age in years *		8.9 (5.65)	12.5 (8.66)
Sex			
	Female	39 (57.4%)	68 (52.7%)
	Male	29 (42.6%)	61 (47.3%)
PRO available †		68	129
CN available ‡		20	58

* mean (SD)

† number of patients with Parent Reported Outcomes

‡ number of patients with Clinical Notes

3.3 Data preparation and features selection

3.3.1 PROs variables

From the estimated 1,300 questions and sub-questions in the questionnaires, 201 were retained for the analysis, based on the estimated relevance by a medical doctor. A large part of the discarded questions regarded details about a specific medical condition, such as the age at wich it appeared. While this could be a useful information to record, it was often not answered and there was a large number of missing values for these questions.

From the 201 questions selected in the PROs, 94 were discarded because they contained more than 97% of the same modality as these questions are not balanced enough and could not yield accurate results.

In total, 107 questions from the PROs were retained for the analysis.

3.3.2 Clinical Notes variables

The SNOMED CT codes availables frome the Clinical Notes contained more than 13,000 codes. After selecting only the codes from the "Findings" section, 4923 codes remained.

The ratio features/observations (63.1) made the analysis extremely challenging.

Aggregating the variables based on the length of their path in the SNOMED CT's tree allowed us to obtain a more reasonable number of features to analyze. However the total number of variables remained high (Table 4) and capping the path length lead to a global loss of granularity. In order to explore the potential associations without relying on a predefined cutoff for this aggregation, we explored the results of the different models for every path length.

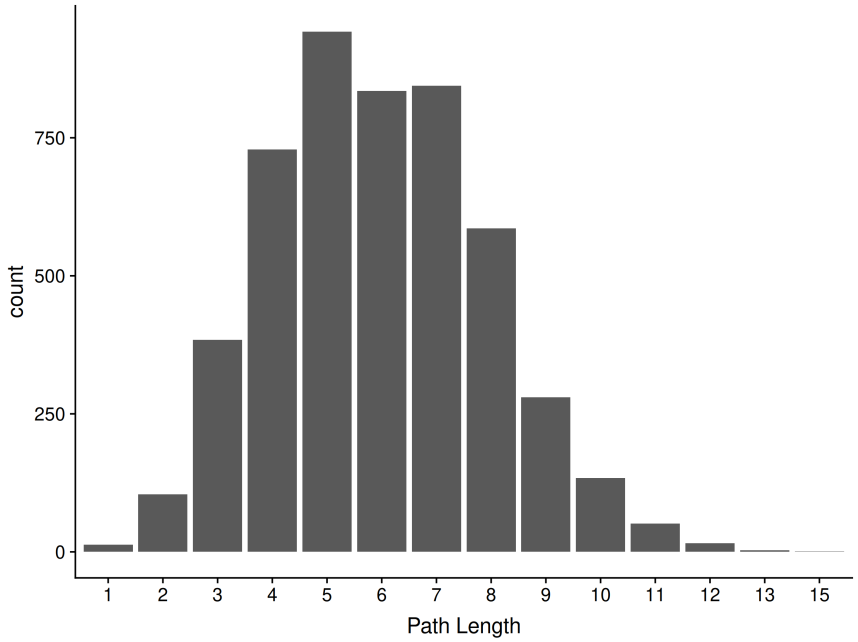

Figure 1: Repartition of the path lengths of the SNOMED CT codes present in the Clinical Notes. The path length represents the depth in the hierarchical tree of the ontology.

3.4 Evaluation of the Clinical Notes processing

The terms chosen for the Clinical notes evaluation and their corresponding precision, recall, and F1 are reported on **table 2** .

The identification of developmental regression had a precision, recall and F1 score of 0.91, 0.78 and 0.84 respectively.

Overall, the performance of the process was satisfactory with a precision, recall and F1 score of 0.89, 0.70 and 0.78 respectively.

Table 2: Evaluation of the processing of the clinical notes by optical character recognition and natural language processing: for each term we compare the results of the automatic extraction to the Gold Standard created by reading the PDFs.

Term	Precision	Recall	F1
Hypotonia	0.97	0.95	0.96
PICA	1.00	0.82	0.90
Gastroesophageal Reflux Disease	0.92	0.80	0.86
Developmental Regression	0.91	0.78	0.84
Lymphedema	0.93	0.74	0.82
Flat Feet	0.74	0.85	0.79
Sleep disorders	0.65	0.79	0.71
Asthma	0.96	0.52	0.68
Ptosis	0.92	0.50	0.65
Cardiac Murmur	0.96	0.46	0.62
Strabismus	0.78	0.47	0.58
Overall *	**0.89**	**0.70**	**0.78**

* macro-averaged values over the different terms

3.5 Determination of the presence of a developmental regression

In the PROs, out of the 197 patients with at least one answer to at least one of the questions regarding developmental regression or having medical records asserting presence of a developmental regression, 113 / 197 (57.4%) were reported as having presented a regression of at least one skill.

in the Clinical notes, 40 / 78 (51.3%) patients had a record of a developmental regression. 16 of these patients had not been identified by the PROs.

3.6 Statistical Analysis

3.6.1 Clinical Notes

- PheWAS

  The different PheWASs realized showed consistent results over the different path lengths. The manhattan plot for path lengths 3, 5, 8 and 15 are displayed in **Figure 2**. When we limit the path length to one, the lack of granularity did not yield relevant results, since

the meaning of the variables is extremely general.

For path lengths of two or three, we obtain one single significant variable (respectively ’Finding by site/Central nervous system finding’, adjusted p-value: 0.0003 and ’Finding by site/Central nervous system finding/Finding of brain’, adjusted p-value: 0.0022). No other feature shows an association with developmental regression. When we go deeper in the tree, we can follow the path of the variable that was associated with the regressions, and "Finding by site/Central nervous system finding/Finding of brain/Epileptic seizure" stays the most significant variable, despite not reaching the significance threshold.

The p-value and odds ratio for the ten most significant variables of the analysis performed without aggregation are shown in **table 3**.

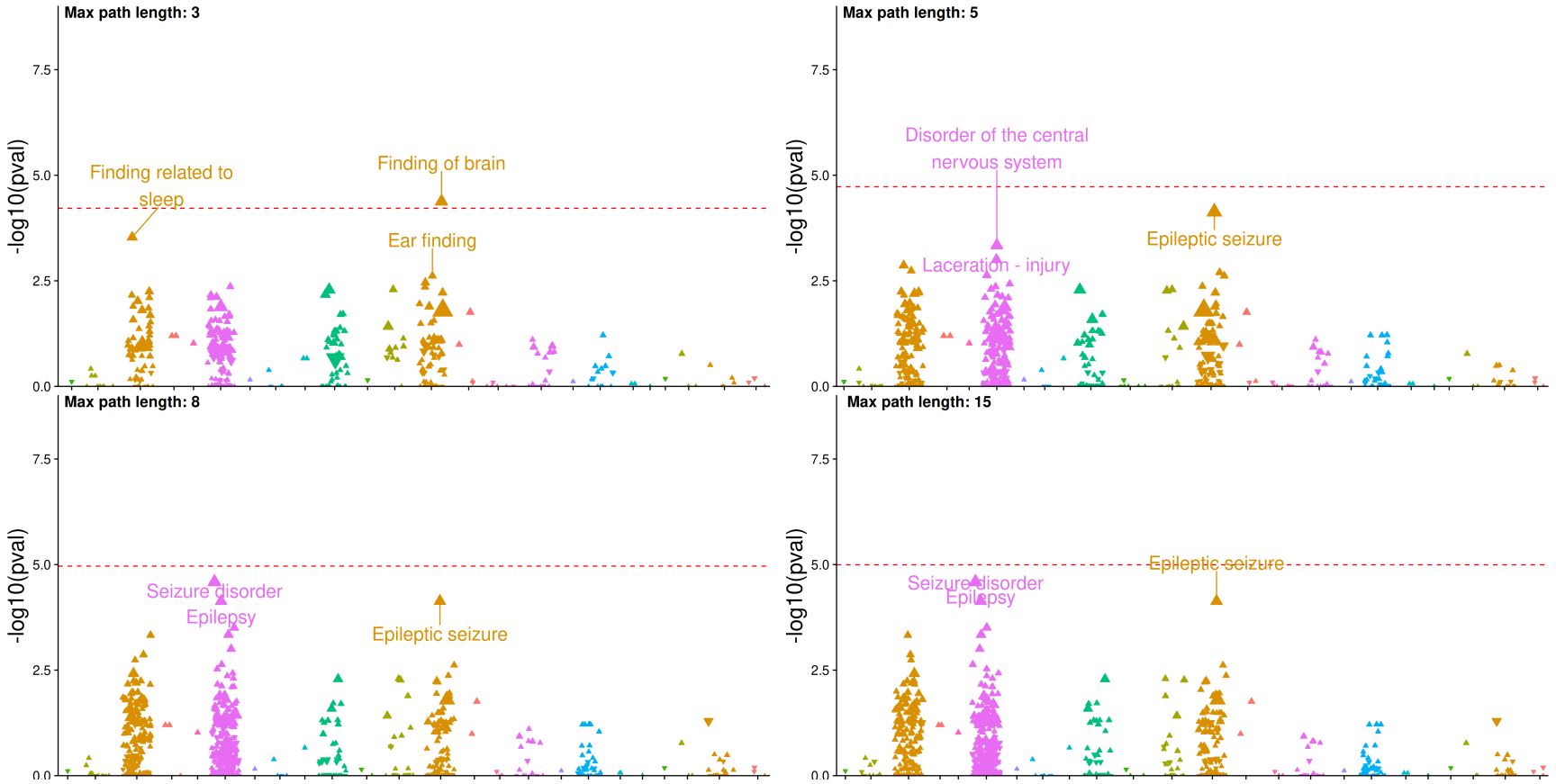

Figure 2: Manhattan plots for the PheWAS on the Clinical notes for paths 3, 5, 8, 15. The plots for all path lengths are shown, by increasing maximum length, from left to right and from top to bottom. The most significant features have been labelled. Each plot follows the same representation design as the Figure X.

Table 3: PheWAS of the Clinical Notes without variable aggregation: 10 most significant results. The concepts related to epileptic seizures are close to the significance threshold after correction for the multiple testing for 4922 variables.

Feature	FDR-adjusted p-value	OR (CI 95% [*])
Seizure disorder	0.098	19.62 (5.58 ; 93.96)
Epilepsy	0.098	16.24 (4.68 ; 76.96)
Epileptic seizure	0.098	16.24 (4.68 ; 76.96)
Disorder of brain	0.317	10.6 (3.12 ; 40.81)
Awake	0.317	8.08 (2.7 ; 26.96)
Disorder of the central nervous system	0.379	11.05 (3.04 ; 47.2)
Laceration - injury	0.771	9.27 (2.74 ; 43.06)
Finding related to ability to perform telephone activities	0.836	5.84 (2 ; 18.41)
Disability	0.836	7.09 (2.14 ; 25.38)
Lymphadenopathy	0.836	7.05 (2.25 ; 27.16)

[*] 95% confidence interval

- Elastic Net

  The Elastic Net produced unstable results up to a path length of 5 in the SNOMED CT tree, where no features were excluded by the regularisation (on average 483.40 (SD = 621.03) features included in the model). The average AUC was 0.73 (SD = 0.05) .

  For the path length 8 to 15, the average AUC was 0.87 (SD = 0.01) . There was on average 6.88 (SD = 2.36) features included in the models.

  The plot of the regularisation path for path length 7, 8, 14, 15 is displayed in **figure 3**, along with the selected features for each path length. "Seizure Disorder", "Epileptic Seizure", "Regression - mental defense mechanism" are consistently retrieved in the model from path length 7.

  The ROC for the Elastic Net modelisations for every path lengths are shown in **figure 6**.

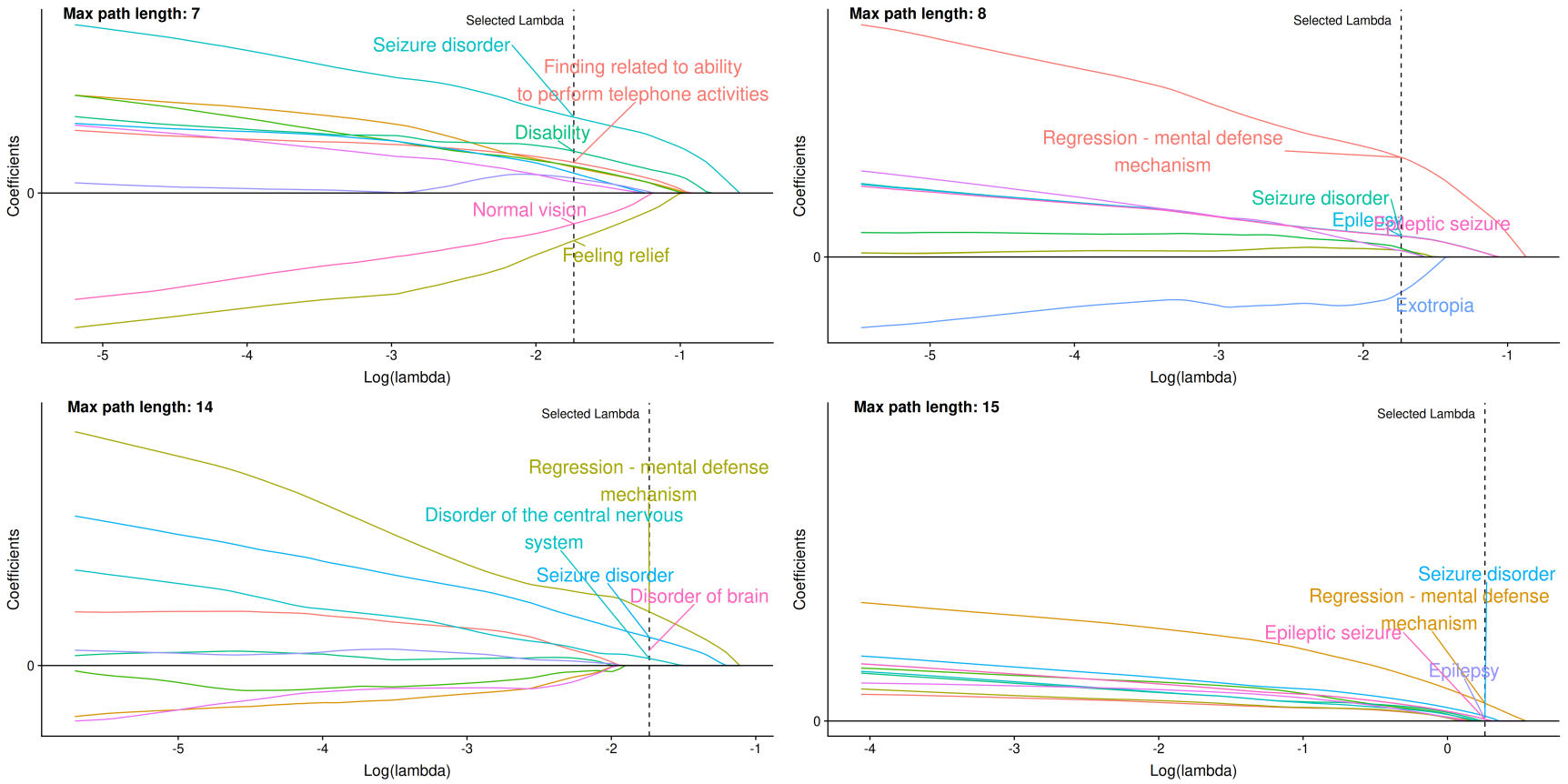

Figure 3: Regularization path of the Elastic Net modelisation for paths 7, 8, 14, 15. Each plot represents the evolution of the values of the different coeficients of the features for a given lambda value. The y axis represents the coefficients of the logistic regression, the x axis shows different lambda values used for the penalization. The dashed vertical line shows the selected lambda value after the parameter search. the higher the lambda, the stronger the penalization, resulting in more features being excluded from the model. The features selected for the best performing lambda value are annotated.

- Random Forest

  The random forest models had an average AUC of 0.71 (SD = 0.08) between path length 1 and 6, increasing to 0.82 (SD = 0.00) for path 8 to 15.

  From path 1 to 6, there was on average 7.00 (SD = 2.61) features with a relative importance superior to 25% of the most important variable, and 2.50 (SD = 1.05) features with a relative importance superior to 50%.

  From path 7 to 15, there was on average 4.00 (SD = 0.50) features with a relative importance superior to 25% of the most important variable, and 2.22 (SD = 0.97) features with a relative importance superior to 50%.

  From path length 3 to 15 the features "Central nervous system finding", "Finding of brain", "Epileptic Seizure" are found consistently as the first or second most important feature of the models.

  The ROC for the Random Forest modelisations for every path lengths are shown in **figure 6**.

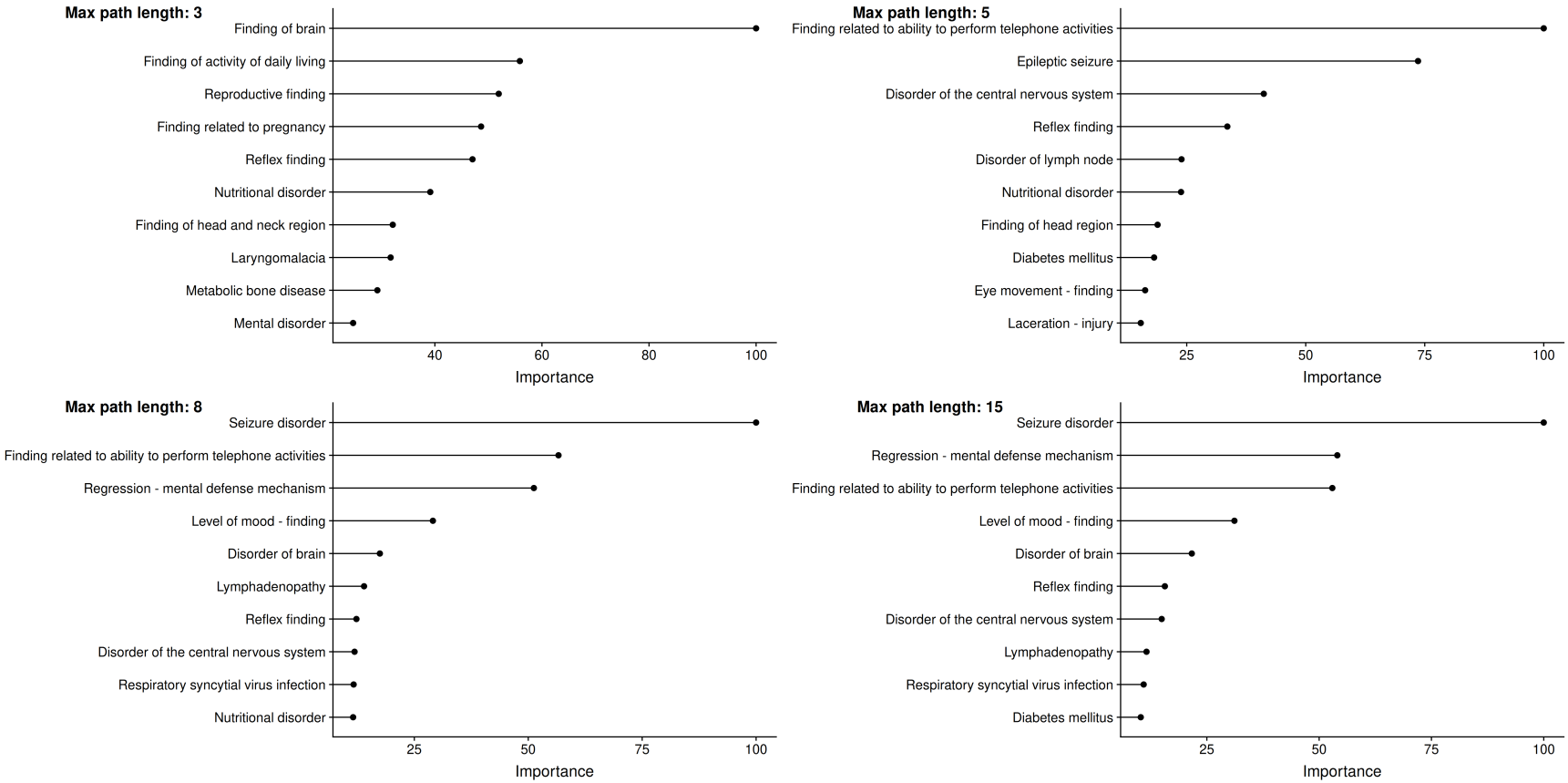

Figure 4: Feature importance of the Random Forest for paths 2, 4, 8, 15. The x axis represents the importance, relative to the most important variable. The ten most important features for each path length are represented.

- Gradient Boosting Machine

  The XGBoost models had an average AUC of 0.70 (SD = 0.09) between path length 1 and 6, increasing to 0.88 (SD = 0.02) for path 7 to 15.

  From path 1 to 6, there was on average 2.67 (SD = 2.73) features with a relative importance superior to 25% of the most important variable, and 1.17 (SD = 0.41) with a relative importance superior to 50%.

  From path 7 to 15, there was on average 2.00 (SD = 0.00) features with a relative importance superior to 25% of the most important variable, and 1.67 (SD = 0.50) with a relative importance superior to 50%.

  Similarily to the Random Forest models, from path length 3 to 15 the features "Central nervous system finding", "Finding of brain", "Epileptic Seizure" are found consistently as the first or second most important feature of the models.

  The ROC for the Gradient Boosting modelisations for every path lengths are shown in **figure 6**.

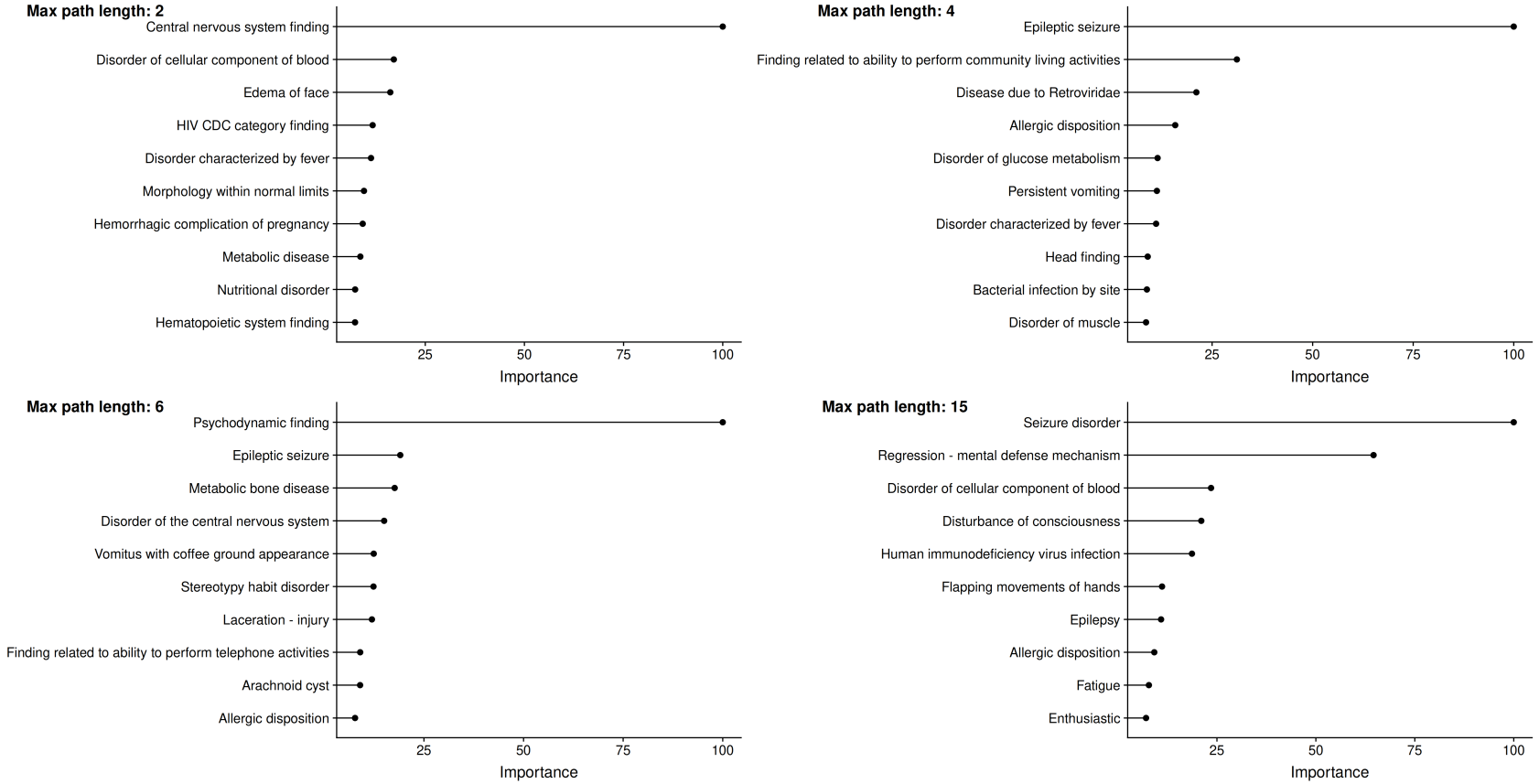

Figure 5: Feature importance of the XGBoost for paths 2, 4, 8, 15. The x axis represents the importance, relative to the most important variable. The ten most important features for each path length are represented in these plot.

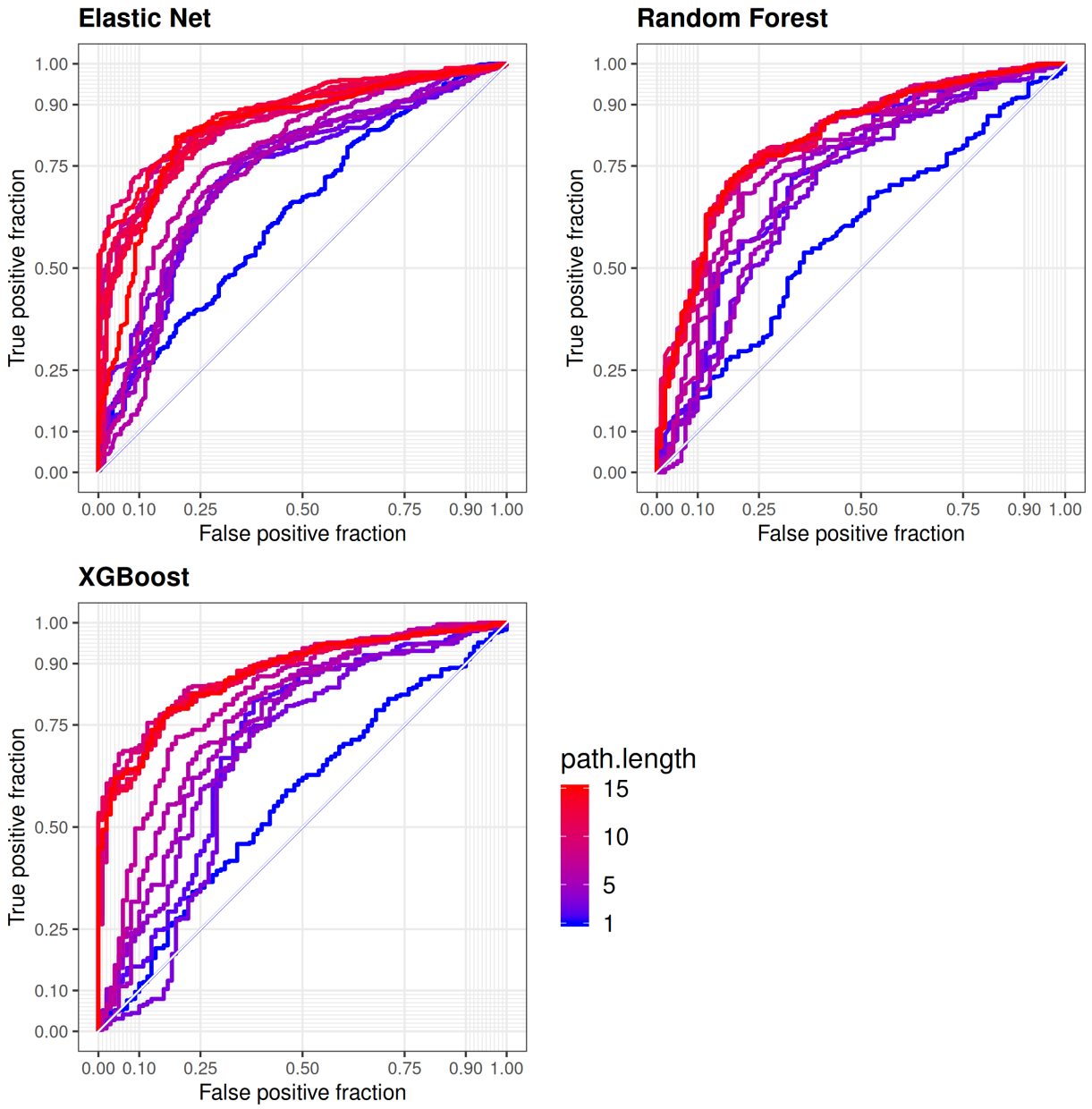

Figure 6: Receiver Operating Characteristics (ROC) curve for the Elastic net, Random forest, and Gradient Boosting Machine models. The curves are reported for every path length.

3.6.2 Parent Reported Outcomes

- Missing values

  The average response rate to the questions in the PRO was 0.71 (SD = 0.32).

  59/107 (55%) questions in the PRO have an answer rate superior to 85%, and 33/107 (31%) questions in the PRO had less than 50% respondents.

- PheWAS

  The PheWAS performed on the PROs did not reveal any significant feature after correction by multiple testing, either using the Bonferroni correction or the Benjamini & Hochberg (FDR) correction (**Table 3**). The most significant feature is the question relating to the presence of aspirations (FDR-corrected p-value: 0.64, OR: 3.52 (1.34 ; 11.15)).

  The manhattan plot is represented on **Figure 7**.

Table 4: PheWAS of the Parent Reported Outcomes: 10 most significant results.

Question	p-value	FDR	OR (CI95%*)
Aspiration	0.018	0.64	3.52 (1.34 ; 11.15)
Swallowing difficulty	0.020	0.64	2.51 (1.19 ; 5.64)
Frequent vomiting	0.020	0.64	2.89 (1.24 ; 7.6)
Holding Breath	0.025	0.64	5.66 (1.52 ; 36.92)
Stools with partially digested food	0.032	0.64	2.34 (1.1 ; 5.28)
Excessive Sweating In Hands And|Or Feet	0.037	0.64	8.98 (1.71 ; 165.94)
Failure To Progress	0.050	0.64	3.71 (1.13 ; 16.93)
Excessive gas	0.055	0.64	3.07 (1.07 ; 11.17)
Two Or More Months On Antibiotics With Little Effect	0.057	0.64	7.43 (1.41 ; 137.1)
Lack Of Perspiration	0.061	0.64	2.00 (0.98 ; 4.25)

* 95% confidence interval

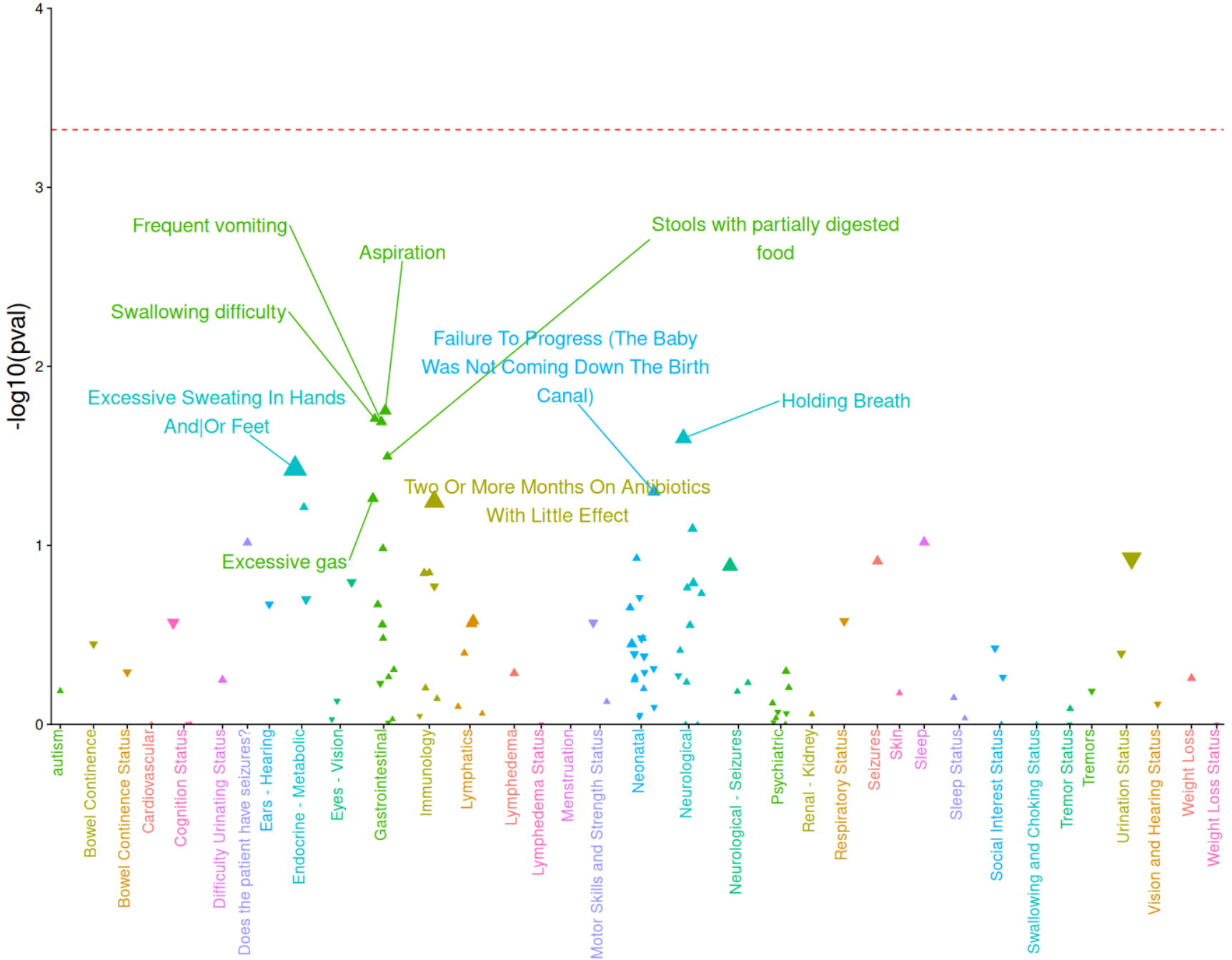

Figure 7: Manhattan plot for the PheWAS on the Parent Reported Outcomes. On the x axis are the different categories from the questionnaires. On the y axis, an inverted logarithmic scale of the p-value. Each triangle represents a question in the PRO. The colors identify the category to which the question belongs. The size of the triangle represents the magnitude of the odds ratio relative to 1. Upwards pointing triangles represent an odds-ratio > 1. The significance threshold line represents the Bonferroni corrected 0.05 p-value. We annotated the five percent most significant questions with an OR > 2, and questions with an OR > 5.

3.7 Exploration of the results: review of the Clinical Notes and PRO

3.7.1 Reading the PDFs

Exploration of the results of the statistical analysis

The analysis of the clinical notes indicated that the presence of epilepic seizures might be associated to the presence of developmental regressions in the Phelan-McDermid syndrome, as this feature was found consistently in the different models and path length. In order to have a deeper insight of the correctness of this association, we manually reviewed the clinical notes of the patients and annotated the presence of epilepsy and developmental regression, as well as wether the developmental regression occured before or after the first clinical manifestations of epilepsy.

Among the patients with a developmental regression, 13/30 (43.3%) patients had a record of epilepsy in their clinical notes. 12/48 (25.0%) patients without a developmental regression were found to have epilepsy. The chi-squared test yielded a p-value of 0.15 for the association between epilepsy and developmental regression.

Among the 13 patients having presented both epilepsy and developmental regression, 7 were found to have a record of epilepsy preceding the regression, and 6 had the first record of epilepsy after the onset of regression.

3.7.2 PRO questions review

One of the questions in the PROs asked specifically about the co-occurence of seizures and developmental regression ("Did onset of seizures correlate with a loss of skill?").
This question also gathered information about the type of skill that was lost, if any.
Among the 34 respondents with seizures, 14 / 34 (41.2%) reported that the onset of seizures correlated with a loss of skill. For 12 / 34 (35.3%) patients, loss of motor skill was reported. A loss of speech abilities has been reported for 6 / 34 (17.6%) patients, and 5 / 34 (14.7%) reported the loss of a complex skills.

4 Discussion

Developmental regressions are a frequent and severe complication in the Phelan-McDermid syndrome, strongly impairing the quality of life of the patients and leading to potentially life threatening medical complications. The mechanisms of these regressions is poorly understood, and no risk factor have been clearly identified. Identifying clinical conditions associated to the presence of a developmental regression could help in characterizating risk factors of developmental regressions in PMS, and better understanding their underlying mechanisms.

In this study we used the Phelan-McDermid Data Network to identify medical conditions associated to developmental regressions, by using statistical analysis and machine learning algorithms to explore the structured data from Parent Reported Outcomes and Clinical Notes, and exploring the most important results by reviewing the unprocessed Clinical Notes and studying the relevant questions in the Parent Reported Outcomes.

The presence of epileptic seizures was the clearest result obtained from the analysis of the Clinical Notes

This medical condition is the only one showing a potential association with developmental regressions in PMS. The significance is not strong enough to overcome the Bonferroni correction except for one variable for the path lengths 2 and 3 (respectively "Finding by site/Central nervous system finding", with a corrected p-value of 0.00384 and "Finding by site/Central nervous system finding/Finding of brain", with a correctedd p-value of 0.0281). However, the variable "Disease/Disorder by body site/Disorder of body system/Disorder of nervous system/Disorder of the central nervous system/Disorder of brain/Seizure disorder" has a p-value of 0.0000217 before correction and would therefore still be inferior to 0.05 if we had 2304 variables (instead of the 4924 in the analysis). Given our choice of the SNOMED CT for it's high number of variables and since we did not have a perfect way of aggregating the variables to reduce their number, we consider the presence of epilepsy as an interesting result for a potential association with developmental regression in PMS.

Reierson et al.[2] studied the association between epilepsy, EEG abnormalities and developmen-

tal regressions in the Phelan-McDermid syndrome on 50 patients, but did not find a significant association between seizures or the fact of presenting an abnormal EEG and developmental regression ($p \geqslant 0.223$).

To our knowledge no other study analyzed the determinants of developmental regression in the Phelan-McDermid syndrome.

Our study utilized the informations available in an international registry. Although this allowed us to gather data for an important number of patients comparing to previous studies on PMS, we did not have a full control on the exact data collected data nor on the completeness of the data.

Firstly, the analysis of the Parent Reported Outcomes was challenging due to the low answer rate, and the analysis of PRO should be interpreted with care as they are prone to memory bias as well as sampling bias.

Secondly, the analysis of the Clinical Notes relied on a preprocessing by optical character recognition and natural language processing. Since we had access to the full Clinical Notes in the database, we could evaluate the output of the processing to confirm that the data reflected the content of the raw text .

The worst performances were observed in terms for which medical doctors tend to not use a specific term (Sleep disorders), or medical conditions that are frequent and commonly reported as a negative sign in the clinical notes, due to the difficulty to interpret negative forms in some sentences (Cardiac Murmur, Asthma, Strabismus, Ptosis). Terms related to rare conditions and having an unambiguous denomination were well detected (Gastroesophageal Reflux Disease, Flat Feet, Hypotonia, PICA, Lymphedema).

Although this evaluation of the processed clinical note showed that it was reliable enough, it remained imperfect. An example of this limitation is the presence of "Regression - mental defense mechanism" in the most important variables of some analyses. The review of the sentences associated to the detection of this variable showed that is was a false positive, triggered by the presence of the word "regression". The patients were marked with this psychiatric concept in

plus of the developmental one.

Finally, analyzing medical codes from the SNOMED CT caused our dataset to contain a very large amount of variables, and to contain redundant features due to the structure of the SNOMED CT terminology. We tried to limit the impact of the large number of variables by utilizing the tree structure of the terminology, but this aggregation method did not allow us to achieve better performance or to eliminate the redundancy of features.

An example of this limitation is the presence of "Seizure Disorder" and "Epilepsy" from the "Disease" section of the SNOMED CT as well as "Epileptic Seizure" from the "Finding by site" category. These variables have similar p-values and have a very similar meaning. This redundancy results from the fact that the hierarchy of the SNOMED CT is not designed for aggregation, and some concepts that could be considered as similar on a medical perspective are located in different branches of the tree. Given the high number of variables in the SNOMED CT and the absence of a pre-curated aggregation method for this ontology, we could not use a different aggregation method in the scope of this work.

Despite these limitations, the results of the different analyses on the clinical notes were consistent and all indicated that epileptic seizures might be associated with developmental regressions in the phelan-McDermid syndrome, without consistently retrieving any other associated condition. Although the manual review of the Clinical Notes did not show a significant association between seizures and developmental regression, the p-value was low.

5 Conclusion

In this article we explored an international registry to find medical conditions associated with developmental regression in the Phelan-McDermid syndrome. The presence of epileptic seizures seems to be associated to this comorbidity, and we did not find any other medical condition possibly associated with developmental regression.

Future prospective studies should further investigate the association of epileptic seizures and

developmental regression in order to assert it and explain the exact relationship between these two conditions in the Phelan-McDermid syndrome.

References

[1] M. C. Phelan, R. C. Rogers, R. A. Saul, G. A. Stapleton, K. Sweet, H. McDermid, S. R. Shaw, J. Claytor, J. Willis, and D. P. Kelly, "22q13 deletion syndrome.," *Am J Med Genet*, Jun 2001.

[2] G. Reierson, J. Bernstein, W. Froehlich-Santino, A. Urban, C. Purmann, S. Berquist, J. Jordan, R. O'Hara, and J. Hallmayer, "Characterizing regression in phelan mcdermid syndrome (22q13 deletion syndrome).," *J Psychiatr Res*, Aug 2017.

[3] N. B. Al Backer, "Developmental regression in autism spectrum disorder.," *Sudan J Paediatr*, 2015.

[4] B. Hagberg, J. Aicardi, K. Dias, and O. Ramos, "A progressive syndrome of autism, dementia, ataxia, and loss of purposeful hand use in girls: Rett's syndrome: report of 35 cases.," *Ann Neurol*, Oct 1983.

[5] G. Baird, T. Charman, A. Pickles, S. Chandler, T. Loucas, D. Meldrum, I. Carcani-Rathwell, D. Serkana, and E. Simonoff, "Regression, developmental trajectory and associated problems in disorders in the autism spectrum: the snap study.," *J Autism Dev Disord*, Nov 2008.

[6] R. Tuchman, "Autism and epilepsy: what has regression got to do with it?," *Epilepsy Curr*, Jul-Aug 2006.

[7] S. Wilson, A. Djukic, S. Shinnar, C. Dharmani, and I. Rapin, "Clinical characteristics of language regression in children.," *Dev Med Child Neurol*, Aug 2003.

[8] V. Hughes, "Scientists track adult regression in autism-related syndrome." Jul 2012.

[9] M. H. Willemsen, J. H. M. Rensen, H. M. J. van Schrojenstein-Lantman de Valk, B. C. J. Hamel, and T. Kleefstra, "Adult phenotypes in angelman- and rett-like syndromes.," *Mol Syndromol*, Apr 2012.

[10] M. G. Figura, A. Coppola, M. Bottitta, G. Calabrese, L. Grillo, D. Luciano, L. Del Gaudio, C. Torniero, S. Striano, and M. Elia, "Seizures and eeg pattern in the 22q13.3 deletion syndrome: clinical report of six italian cases.," *Seizure*, Oct 2014.

[11] D. Zhang, F. Bedogni, S. Boterberg, C. Camfield, P. Camfield, T. Charman, L. Curfs, C. Einspieler, G. Esposito, B. De Filippis, R. P. Goin-Kochel, G. U. Hoglinger, D. Holzinger, A.-M. Iosif, G. E. Lancioni, N. Landsberger, G. Laviola, E. M. Marco, M. Muller, J. L. Neul, K. Nielsen-Saines, A. Nordahl-Hansen, M. F. O'Reilly, S. Ozonoff, L. Poustka, H. Roeyers, M. Rankovic, J. Sigafoos, K. Tammimies, G. S. Townend, L. Zwaigenbaum, M. Zweckstetter, S. Bolte, and P. B. Marschik, "Towards a consensus on developmental regression.," *Neurosci Biobehav Rev*, Aug 2019.

[12] C. Lord, M. Rutter, and A. Le Couteur, "Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders.," *J Autism Dev Disord*, Oct 1994.

[13] S. U. Dhar, D. del Gaudio, J. R. German, S. U. Peters, Z. Ou, P. I. Bader, J. S. Berg, M. Blazo, C. W. Brown, B. H. Graham, T. A. Grebe, S. Lalani, M. Irons, S. Sparagana, M. Williams, J. A. r. Phillips, A. L. Beaudet, P. Stankiewicz, A. Patel, S. W. Cheung, and T. Sahoo, "22q13.3 deletion syndrome: clinical and molecular analysis using array cgh.," *Am J Med Genet A*, Mar 2010.

[14] L. Soorya, A. Kolevzon, J. Zweifach, T. Lim, Y. Dobry, L. Schwartz, Y. Frank, A. T. Wang, G. Cai, E. Parkhomenko, D. Halpern, D. Grodberg, B. Angarita, J. P. Willner, A. Yang, R. Canitano, W. Chaplin, C. Betancur, and J. D. Buxbaum, "Prospective investigation of

autism and genotype-phenotype correlations in 22q13 deletion syndrome and shank3 deficiency.," *Mol Autism*, Jun 2013.

[15] A. Denayer, H. Van Esch, T. de Ravel, J.-P. Frijns, G. Van Buggenhout, A. Vogels, K. Devriendt, J. Geutjens, P. Thiry, and A. Swillen, "Neuropsychopathology in 7 patients with the 22q13 deletion syndrome: Presence of bipolar disorder and progressive loss of skills.," *Mol Syndromol*, Jun 2012.

[16] M. A. Manning, S. B. Cassidy, C. Clericuzio, A. M. Cherry, S. Schwartz, L. Hudgins, G. M. Enns, and H. E. Hoyme, "Terminal 22q deletion syndrome: a newly recognized cause of speech and language disability in the autism spectrum.," *Pediatrics*, Aug 2004.

[17] M. Macedoni-Luksic, D. Krgovic, B. Zagradisnik, and N. Kokalj-Vokac, "Deletion of the last exon of shank3 gene produces the full phelan-mcdermid phenotype: a case report.," *Gene*, Jul 2013.

[18] S. Serret, S. Thümmler, E. Dor, S. Vesperini, A. Santos, and F. Askenazy, "Lithium as a rescue therapy for regression and catatonia features in two SHANK3 patients with autism spectrum disorder: case reports," *BMC Psychiatry*, vol. 15, p. 107, May 2015.

[19] D. M. Cochoy, A. Kolevzon, Y. Kajiwara, M. Schoen, M. Pascual-Lucas, S. Lurie, J. D. Buxbaum, T. M. Boeckers, and M. J. Schmeisser, "Phenotypic and functional analysis of shank3 stop mutations identified in individuals with asd and/or id.," *Mol Autism*, 2015.

[20] A. Philippe, N. Boddaert, L. Vaivre-Douret, L. Robel, L. Danon-Boileau, V. Malan, M.-C. de Blois, D. Heron, L. Colleaux, B. Golse, M. Zilbovicius, and A. Munnich, "Neurobehavioral profile and brain imaging study of the 22q13.3 deletion syndrome in childhood.," *Pediatrics*, Aug 2008.

[21] R. E. Frye, D. Cox, J. Slattery, M. Tippett, S. Kahler, D. Granpeesheh, S. Damle, A. Legido, and M. J. Goldenthal, "Mitochondrial dysfunction may explain symptom variation in phelan-mcdermid syndrome.," *Sci Rep*, Jan 2016.

[22] S. De Rubeis, P. M. Siper, A. Durkin, J. Weissman, F. Muratet, D. Halpern, M. D. P. Trelles, Y. Frank, R. Lozano, A. T. Wang, J. L. J. Holder, C. Betancur, J. D. Buxbaum, and A. Kolevzon, "Delineation of the genetic and clinical spectrum of phelan-mcdermid syndrome caused by shank3 point mutations.," *Mol Autism*, 2018.

[23] N. B. Al Backer, "Developmental regression in autism spectrum disorder.," *Sudan J Paediatr*, 2015.

[24] T. Deonna, A. L. Ziegler, J. Moura-Serra, and G. Innocenti, "Autistic regression in relation to limbic pathology and epilepsy: report of two cases.," *Dev Med Child Neurol*, Feb 1993.

[25] R. Tuchman, "Autism and epilepsy: what has regression got to do with it?," *Epilepsy Curr*, Jul-Aug 2006.

[26] E. W. Viscidi, E. W. Triche, M. F. Pescosolido, R. L. McLean, R. M. Joseph, S. J. Spence, and E. M. Morrow, "Clinical characteristics of children with autism spectrum disorder and co-occurring epilepsy.," *PLoS One*, 2013.

[27] D. C. Tarquinio, W. Hou, A. Berg, W. E. Kaufmann, J. B. Lane, S. A. Skinner, K. J. Motil, J. L. Neul, A. K. Percy, and D. G. Glaze, "Longitudinal course of epilepsy in rett syndrome and related disorders.," *Brain*, Feb 2017.

[28] S. E. Daugherty, S. Wahba, and R. Fleurence, "Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research.," *J Am Med Inform Assoc*, Jul-Aug 2014.

[29] C. Kothari, M. Wack, C. Hassen-Khodja, S. Finan, G. Savova, M. O'Boyle, G. Bliss, A. Cornell, E. J. Horn, R. Davis, J. Jacobs, I. Kohane, and P. Avillach, "Phelan-mcdermid syndrome data network: Integrating patient reported outcomes with clinical notes and curated genetic reports.," *Am J Med Genet B Neuropsychiatr Genet*, Oct 2018.

[30] "Health information privacy." https://www.hhs.gov/hipaa/index.html, Aug. 2015. Accessed: 2018-6-21.

[31] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications.," *J Am Med Inform Assoc*, Sep-Oct 2010.

[32] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology.," *Nucleic Acids Res*, Jan 2004.

[33] I. Rombach, O. Rivero-Arias, A. M. Gray, C. Jenkinson, and O. Burke, "The current practice of handling and reporting missing outcome data in eight widely used proms in rct publications: a review of the current literature.," *Qual Life Res*, Jul 2016.

[34] I. M. Johnstone and D. M. Titterington, "Statistical challenges of high-dimensional data.," *Philos Trans A Math Phys Eng Sci*, Nov 2009.

[35] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford, "Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations.," *Bioinformatics*, May 2010.

[36] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, ACM, 2016.

Abstract

introduction: Developmental regression is a frequent and severe complication in the Phelan-McDermid Syndrome (PMS). Little is known about the causes and mechanisms of this condition, and no treatment is available to date. Finding medical conditions associated to the apparition of developmental regressions in PMS could help to understand the mechanisms of the regressions.

materials and methods: Using the Phelan-McDermid Data Network allowed us to analyze the Parent Reported Outcomes (PRO) from 233 patients and the full clinical notes history processed by natural language processing (NLP) of 78 patients with PMS, expressed as SNOMED CT codes.

We analyzed the PRO with a Phenome-Wide Association Study (PheWAS).

After ascertaining the quality of the NLP processing by comparing the codes to a set of manually annotated conditions, clinical notes were analyzed with a PheWAS and several machine learning modelisations: Elastic Net logistic regression, Random Forest, Gradient Boosting Machine. We attempted to aggregate the SNOMED CT by limiting the depth of the tree to reduce the number of features. We identified important variables to predict the presence of developmental regressions.

We then manually reviewed the clinical notes and the PRO to fully analyze the available information for these features.

results: The PRO analysis did not yield significant features. In the clinical notes, seizure disorders were close to significance in all the PheWAS (FDR $= 0.098$). it was also consistently found as the most important factor in the machine learning modelisations. When reading of the clinical notes, 13/30 (43.3%) of the patients with regression also had epilepsy, and 12/48 (25.0%) without regression had epilepsy.

conclusion: This study found a consistent link between epileptic seizures and developmental regressions in PMS. Further studies should confirm it, investigate it's nature and potential mechanisms. We did not identify other conditions associated to developmental regressions.

keywords: Phelan-McDermid Syndrome, 22q13.3 deletion syndrome, developmental regression, Clinical Datawarehouse