



Université Paris Cité

École doctorale de Sciences Mathématiques de Paris Centre (ED 386)
Centre de Recherche des Cordeliers UMRS1138 Inserm, Université Paris Cité

Extraction et prédition des réponses aux chimiothérapies à partir des données hospitalières

Thèse de doctorat en INFORMATIQUE

Présentée par ALICE ROGIER

Dirigée par Adrien Coulet et Bastien Rance

Soutenue le 17 octobre 2024

Membres du jury :

PR. FLEUR MOUGIN	UNIVERSITÉ DE BORDEAUX	Rapportrice
PR. BRIGITTE SÉROUSSI	SORBONNE UNIVERSITÉ	Rapportrice
PR. SANDRA BRINGAY	UNIVERSITÉ PAUL VALÉRY DE MONTPELLIER	Examinateuse
PR. MARIE-ANNE LORIOT	UNIVERSITÉ PARIS CITÉ	Examinateuse
DR. MAUD TOULMONDE	INSTITUT BERGONIÉ À BORDEAUX	Examinateuse
DR. VIANNEY JOUHET	UNIVERSITÉ DE BORDEAUX	Invité
DR. ADRIEN COULET	INRIA PARIS	Co-directeur
DR. BASTIEN RANCE	UNIVERSITÉ PARIS CITÉ	Co-directeur

TABLE DES MATIÈRES

Table des matières	iv
Liste des figures	ix
Liste des tableaux	xii
Remerciements	1
Résumé	3
Abstract	5
Introduction	9
I Contexte	13
1 Les chimiothérapies à l'heure des entrepôts de données	15
1.1 Le principe des chimiothérapies	16
1.1.1 Le cancer et sa progression	16
1.1.2 La place des chimiothérapies dans le traitement du cancer	17
1.1.3 L'action des molécules anticancéreuses des chimiothérapies	18
1.1.4 Les schémas thérapeutiques de chimiothérapies	19
1.2 Les chimiothérapies dans la pratique	20
1.2.1 Les réponses patients-dépendantes des chimiothérapies	20
1.2.2 Le déroulement des chimiothérapies	22
1.2.3 Le suivi des réponses aux chimiothérapies	23
1.2.4 L'adaptation des chimiothérapies aux réponses des patients	24
1.2.5 L'adhésion aux chimiothérapies	24
1.2.6 Les adaptations praticien-dépendantes aux protocoles	25
1.3 Évaluation des réponses aux chimiothérapies dans les études cliniques	26
1.3.1 Évaluation des schémas thérapeutiques	26
1.3.2 Évaluation de l'adhésion aux schémas thérapeutiques	26
1.3.3 La nécessité d'un consensus sur les schémas thérapeutiques et le potentiel des DPI	27
1.4 Les entrepôts de données de santé	28
1.4.1 Les données de santé et les Dossiers Patients Informatisés	28
1.4.2 La provenance et l'hétérogénéité des données des DPI	28
1.4.3 L'alimentation des entrepôts de données hospitaliers	29
1.4.4 Le schéma en étoile d'i2b2	29
1.4.5 Les entrepôts de données de l'HEGP et du CHU de Bordeaux	29
1.4.6 Les enregistrements sur les chimiothérapies	30
1.4.7 Les enregistrements de réponses aux chimiothérapies	30
1.4.8 Les données de décès de l'Insee	32
1.5 Les défis à relever pour réutiliser les DPI à des fins de recherche	32

1.5.1	Données dédiées à la recherche <i>vs.</i> données de vie réelles réutilisées pour la recherche	32
1.5.2	La qualité des DPI et leur validité	32
1.5.3	La nécessité d'extraire, de représenter et d'intégrer les informations des DPI	32
1.5.4	Différence de conclusion et implications des résultats obtenus avec des DPI	33
1.6	Bilan : Les défis de la réutilisation des DPI pour étudier les réponses aux chimiothérapies	33
II	État de l'art	35
2	Représentation de connaissances	37
2.1	Donnée, information, connaissance	38
2.2	Représentation de connaissances	39
2.2.1	Les représentations des connaissances dans le domaine bio-médical	39
2.2.2	Bases des logiques de descriptions	43
2.2.3	Le raisonnement automatique	44
2.2.4	Ontologies, bases de connaissances, système à base de connaissances, graphes de connaissances	44
2.3	Représenter les connaissances avec les outils du Web sémantique	46
2.3.1	Les outils du Web sémantique pour implémenter la représentation d'un domaine dans la pratique	46
2.3.2	Resource Description Framework (RDF)	47
2.3.3	Identifier les ressources sur le Web avec les URI	48
2.3.4	Des modèles de référence à réutiliser	49
2.3.5	Gérer et requêter un graphe de connaissances avec un triple store et SPARQL	51
2.4	Raisonnement et fouille de graphe	52
2.4.1	Le raisonnement dans un système à base de connaissances	52
2.4.2	Vers la fouille des graphes de connaissances	54
2.5	État de l'art des modèles de données spécifiques aux chimiothérapies	55
2.5.1	Note sur les représentations temporelles de la TEO et de la TO	57
2.6	Bilan : pourquoi les ontologies ?	58
3	Extraction de connaissances et prédiction d'événements	59
3.1	Extraction de connaissances	60
3.1.1	Extraction de connaissances à partir du texte libre	60
3.1.2	Des mesures pour extraire des connaissances	63
3.2	Prédiction d'événements	67
3.2.1	Analyse de survie classique	67
3.2.2	Régression logistique pour l'analyse de survie	81
3.2.3	Comparaison des modèles de survie étudiés	84
3.2.4	Prédictions d'événements avec des modèles de survie	86
3.3	Bilan : quels outils pour extraire et prédire les réponses aux chimiothérapies ?	91

III Contributions 93

4 Représenter les chimiothérapies et leurs réponses avec ChemoOnto et OntoTox	95
4.1 Représenter les chimiothérapies et leurs réponses avec des ontologies	97
4.1.1 Pour standardiser les concepts relatifs aux chimiothérapies et leurs réponses	97
4.1.2 Pour les connecter à des modèles de connaissances de référence	100
4.1.3 Discussion sur les choix de représentation	102
4.2 Extraire les connaissances sur les chimiothérapies et leurs réponses depuis les entrepôt de données	106
4.2.1 Extraire les connaissances sur les chimiothérapies	106
4.2.2 Extraire les connaissances sur la survenue de toxicités	114
4.2.3 Discussion sur les méthodes d'extraction d'informations	116
4.3 Utiliser les ontologies instanciées comme des bases de connaissances	119
4.3.1 Instanciation de ChemoOnto	119
4.3.2 Instanciation d'OntoTox	120
4.4 Utiliser les ontologies pour découvrir de nouvelles connaissances	122
4.4.1 Raisonnement temporel	122
4.4.2 Comparer les schémas thérapeutiques avec ChemoKG et les plongements de graphes	125
4.5 Bilan	126
5 ProtoDrift : Mesurer l'adhésion aux chimiothérapies	129
5.1 Motivation pour une nouvelle mesure de l'adhésion aux chimiothérapies	130
5.1.1 Liens entre adhésion et réponses aux chimiothérapies	130
5.1.2 Hypothèses sur la mesure actuelle d'adhésion et objectifs de ProtoDrift	132
5.1.3 Formalisation des concepts de chimiothérapies	132
5.2 Définition de ProtoDrift	134
5.2.1 Dissimilarité administration	134
5.2.2 Dissimilarité médicament	136
5.2.3 Dissimilarité intra-cycle et dissimilarité inter-cycle	138
5.2.4 Dissimilarités : du cycle à la ligne	139
5.2.5 Exemple de calcul de dissimilarité	140
5.3 Optimisation des poids de ProtoDrift	142
5.3.1 Introduction à l'optimisation	142
5.3.2 Objectifs de l'optimisation	143
5.3.3 Méthode d'optimisation	144
5.4 Analyse comparative entre ProtoDrift et la RDI	151
5.4.1 Analyse du pouvoir discriminant	153
5.5 Exploration de l'impact du temps par rapport à la dose sur la survie globale	154
5.6 Bilan sur les méthodes proposées et leurs évaluations	155
5.7 Application de ProtoDrift sur les données l'HEGP et du CHU de Bordeaux	157
5.7.1 Sources de données et conception de l'étude	157
5.7.2 Correspondance entre les noms de localisation de cancer à l'HEGP et au CHU de Bordeaux	162
5.7.3 Évaluation de ProtoDrift sur deux jeux de données indépendants .	163

5.7.4	Configuration des paramètres d'application	163
5.8	Résultats d'application	164
5.8.1	Performances prédictives de la survie globale à 5 ans de la cohorte respiratoire et thoracique en première ligne	164
5.8.2	Évaluation de ProtoDrift	167
5.8.3	Optimisation des poids de ProtoDrift	171
5.8.4	Exploration de l'impact du temps par rapport à la dose sur la survie globale	172
5.9	Discussion et perspectives	173
5.9.1	Interprétation des résultats d'application	173
5.9.2	Optimisation des poids de ProtoDrift	177
5.9.3	Conception de la mesure ProtoDrift	180
5.10	Bilan	181
5.10.1	Résumé des contributions	181
5.10.2	Implications pour la recherche clinique	181
5.10.3	Défis et perspectives	182
5.10.4	Conclusion	182
IV	Conclusion et perspectives	183
Conclusion et perspectives		185
V	Annexes	187
Annexes		189
Bibliographie		217

TABLE DES FIGURES

2.1	Modélisation de la temporalité d'un contrat de travail entre une entreprise et un salarié avec les modèles de réification et 4D-fluent. À gauche le modèle de réification utilisé dans la Time Event Ontology. À droite le modèle 4D-fluent modélisable avec la TO. La figure est tirée de l'article "Temporal representation and reasoning in OWL 2" de BATSAKIS et al. [15].	58
3.1	Reconnaissances des entités d'intérêts maladie et symptôme dans la même phrase.	63
3.2	Liaison des entités d'intérêt avec le parseur de dépendance.	63
3.3	Utilisation des parseurs de dépendance pour l'extraction de relations entre les entités d'intérêt maladie et symptôme. Figures tirés de l'article de HASSAN et al. [84].	63
3.4	Représentation graphique des timelines de patients 1, 2 et 3. Les cercles indiquent une censure tandis que les losanges indiquent la survenue de l'événement. Pour le patient 1, l'événement est survenu à $T_1 = 2.7$ et $d_1 = 1$ indique que l'événement a été observé. Pour le patient 2, le temps observé est $T_2 = 3.4$ et $d_2 = 0$ indique que les données sont censurées à cette date. Pour le patient 3, l'événement est survenu à $T_3 = 5.6$ et $d_3 = 1$ indique que l'événement a été observé. L'image a été adaptée et provient de l'article de SURESH, SEVERN et GHOSH [195].	71
3.5	Figure 3.5a : Courbes de Kaplan-Meier de la survie sans progression (SSP) et de survie globale (SG) dans l'étude de MENG et al. [131], comparant les patients recevant une chimiothérapie seule versus ceux recevant une chimiothérapie avec thérapie ciblée. Les p-valeurs des tests de log-rank montrent que les différences ne sont pas statistiquement significatives pour la SG ($p = 0.93$) et la SSP ($p = 0.29$). Graphiquement, cela est visible car les courbes se croisent. Figure 3.5b : Courbes de Kaplan-Meier de la survie globale (SG) dans l'étude de BREADNER et al. [24], comparant les patients avec une RDI $\leq 80\%$ et ceux avec une RDI $> 80\%$. Les p-valeurs des tests de log-rank montrent une différence statistiquement significative pour la capécitabine ($p = 0.0205$) et l'oxaliplatine ($p = 0.0159$), mais pas pour le 5-FU ($p = 0.3747$). Graphiquement, on observe en effet que les courbes se croisent dans le graphique de l'oxaliplatine, ce qui n'est pas le cas des deux autres graphiques.	75
3.6	Résultats des analyses de régression de Cox : analyses univariées et multivariées des facteurs de risque pour la SG et la SSP dans l'étude de MENG et al. [131].	79
3.7	Résultats des analyses de régression de Cox : analyses multivariées des facteurs de risque pour la SG dans l'étude de BREADNER et al. [24].	80
3.8	Formats des données de survie : temps continu (à gauche) et temps discret (à droite). Le modèle de temps discret utilise le format de droite avec une entrée par intervalle de temps. La figure est tirée de l'article de SURESH, SEVERN et GHOSH [195]	84
3.9	Processus de construction et de test du modèle, illustrant la division des données, l'entraînement du modèle et l'évaluation des performances. Figure tirée de l'article de SURESH, SEVERN et GHOSH [195]	89

4.1	Structure de ChemoOnto, illustrée avec un exemple de son instanciation avec un cycle standard (gauche) et une chimiothérapie suivie par un patient (droite)	98
4.2	Structure d'OntoTox illustrée avec un exemple de son instanciation depuis une extraction du texte libre : "œsophagite de grade II"	101
4.3	Structure des résultats de la requête pour obtenir le déroulement des schémas thérapeutiques théoriques. Les points "..." indiquent que les résultats continuent.	107
4.4	Correspondance entre clé et nom de molécules. Les points "..." indiquent que les résultats continuent.	107
4.5	Représentation ontologique du schéma thérapeutique 1357 avec ChemoOnto. Pour alléger la visualisation, les propriétés de données associées sont indiquées en italique directement dans les instances, plutôt qu'avec des flèches, et elles ne sont indiquées que dans une seule des quatre instances de "Theo-DrugAdministration"	109
4.6	Représentation tabulaire du schéma thérapeutique 1357 avec ChemoOnto. Chaque ligne de la table correspond à une administration d'un schéma thérapeutique spécifique avec ses modalités.	110
4.7	Structure des résultats de la requête pour obtenir le déroulement des lignes suivies de chimiothérapie. Les " " indiquent une valeur identique à la dernière valeur observée de la colonne, tandis que les points "..." indiquent que les résultats continuent.	111
4.8	Représentation ontologique de la première ligne du patient P1 suivant le schéma thérapeutique 1357 avec ChemoOnto. Pour alléger la visualisation, les propriétés de données associées sont indiquées en italique directement dans l'instance, plutôt qu'avec des flèches, et elles ne sont indiquées que dans une seule des huit instances de "DrugAdministration"	113
4.9	Représentation tabulaire de la première ligne du patient P1 suivant le schéma thérapeutique 1357 avec ChemoOnto. Les " " indiquent une valeur identique à la dernière valeur observée de la colonne, tandis que les points "..." indiquent que les résultats continuent.	114
4.10	Lier les entités toxicité et grade grâce au parseur de dépendance. L'entité toxicité est surlignée en orange et l'entité grade est surlignée en bleue	115
4.11	De l'extraction des toxicités à partir de différentes sources de données à leurs instanciation dans OntoTox	116
4.12	Comparaison entre le pourcentage de réduction enregistré dans le champ "réduction" du logiciel Chimio, et le pourcentage de réduction calculée avec les valeurs des administrations théoriques et réelles.	118
4.13	Instanciation d'OntoTox avec une cohorte de patients atteints d'un cancer du poumon.	120
4.14	Diagramme d'intersection des toxicités extraites depuis les différentes sources	121
4.15	Instanciation de ChemoOnto et OntoTox pour 3 923 patients avec des items de questionnaires et des enregistrements dans le logiciel Chimio	123
4.16	Règle SWRL pour inférer la propriété time:inside entre les instances d'instants liés à des toxicités et les instances d'intervalles liés à des cycles suivis	123

4.17	Inférer la propriété time: inside entre les instances d'instants liés à des toxicités et les instances d'intervalles liés à des cycles suivis	124
5.1	La dérive sur un exemple	131
5.2	Représentation scématique des entrées et sorties de l'algorithme d'alignement de ProtoDrift.	137
5.3	Calcul des différentes dissimilarités mesurant la dérive de première ligne suivie par le patient fictif de la figure 5.1.1.	140
5.4	Représentation schématique, sous forme d'arbre horizontal, des dépendances entre les différentes dissimilarités qui composent ProtoDrift. Sur la gauche, les feuilles de l'arbre correspondent aux dissimilarités temporelles et de dose d'administration. Les dissimilarités sont sommées à chaque noeud, correspondant à différentes étapes du traitement, jusqu'à la dissimilarité ligne, qui est la racine de l'arbre, sur la droite du schéma. Chaque dissimilarité au niveau des noeuds est la somme normalisée de ses dissimilarités enfants. $ \mathcal{A}_m $ correspond au maximum entre les administrations de cycles réels et théoriques d'un médicament anticancéreux m . $ \mathcal{M} $ correspond au nombre de médicaments anticancéreux distincts dans le cycle. l correspond au nombre de cycles réels suivis par les patients dans la ligne de traitement.	142
5.5	Structure de la table ProtoDrift de départ, avec une entrée par administration. Tous les poids associés aux dissimilarités sont égaux à 1. Les dissimilarités de cette table correspondent aux dissimilarités du "ProtoDrift Naïf" (NP) ($\omega_d = \omega_t = \omega_{\text{intra}} = \omega_{\text{inter}} = 1 \Leftrightarrow \alpha = \beta = 1/2$). Les " " signifient une valeur identique à l'entrée supérieure. Les ":" signifient que les valeurs peuvent changer. La table est groupée par patient, localisation, ligne de traitement, cycle et médicament. Les dissimilarités sont colorées en fonction du niveau d'arborescence auquel elles sont associées (cf. figure 5.4) : jaune foncé pour la ligne, bleu pour le cycle, vert pour le sous-ensemble de cycle restreint aux administrations d'un même médicament, rose pour l'administration. La couleur verte a été choisie pour le niveau de δ_m car les médicaments, paire (molécule, mode d'administration), sont colorés en jaune clair, et les cycles en bleu.	148
5.6	La table ProtoDrift Naïf (figure 5.5) est filtrée selon une localisation de cancer et un numéro de ligne de traitement, et associée aux données de survie. La table contient donc une entrée par patient.	149
5.7	Étapes 1 et 2 de l'algorithme d'optimisation. Au cours de la première étape, des échantillons <i>bootstrap</i> sont définis à partir des numéros de patients de la table ProtoDrift Naïf filtrée (cf. figure 5.6). Au cours de la deuxième étape, des modèles de régression logistique et de Cox avec $\delta_{\text{line}}(\alpha = \frac{1}{2}, \beta = \frac{1}{2})$ et ADRDI comme variables explicatives (cf. section 5.3.3.3) sont ajustés et prédits, et leurs performances de prédiction sont évaluées. La performance finale des modèles correspond à la moyenne des scores de prédiction obtenus sur les échantillons <i>bootstrap</i> . La réalisation de ces deux étapes permettent de réaliser la double analyse comparative NP-BADRI que nous verrons dans la section 5.4.	150

5.8	Étape 3 de l'algorithme d'optimisation. Pour chaque combinaison (α, β) , nous calculons la dissimilarité $\delta_{line}(\alpha, \beta)$ correspondante. Nous utilisons les échantillons définis lors de la première étape et ajustons et prédisons les modèles de régression logistique et de Cox avec $\delta_{line}(\alpha, \beta)$ comme variable explicative sur les échantillons <i>bootstrap</i> . Cette troisième étape permet de réaliser l'analyse comparative décrite dans la section 5.4 et l'exploration de l'impact relatif des poids sur la survie décrite dans la section 5.5.	151
5.9	Interprétation des profils de <i>heatmap</i>	154
5.10	Rares images d'un ProtoDrift. Gauche : avatar de ProtoDrift . Comme nous le verrons, en général, plus la dérive est forte, plus les chances de survie sont faibles. À droite, avatar de ProtoDrift-Surv , ce dernier paie les conséquences d'un dérapage incontrôlé.	156
5.11	Conception de l'étude. La couleur bleue est associée à l'HEGP et la couleur bordeaux au CHU de Bordeaux. Les cohortes spécifiques à l'HEGP sont donc encadrées en bleu, celles spécifiques au CHU de Bordeaux sont encadrées en bordeaux, et les cohortes communes aux deux hôpitaux sont encadrées en orange.	158
5.12	Caractéristiques démographiques, de survie globale et des métriques statistiques qualitatives des cohortes incluses dans l'analyse (1/2)	160
5.13	Caractéristiques démographiques, de survie globale et des métriques statistiques qualitatives des cohortes incluses dans l'analyse (2/2)	161
5.14	Courbes de Kaplan-Meier construites sur les quartiles de 1–ADRFI (gauche), δ_{line} de NP (milieu) et C-index- δ_{line} (droite) à la fin de la première ligne de la cohorte respiratoire et thoracique (R/T) de l'HEGP.	165
5.15	Tableau de résultats de l'analyse comparative des modèles de régression de Cox pour BADRFI, NP et C-index-OP sur les performances prédictives de la survie globale à 5 ans pour la cohorte de première ligne R/T de l'HEGP .	165
5.16	Courbes de Kaplan-Meier construites sur les quartiles de 1–ADRFI (gauche), δ_{line} de NP (milieu) et C-index- δ_{line} (droite) à la fin de la première ligne de la cohorte respiratoire et thoracique (R/T) du CHU de Bordeaux.	166
5.17	Tableau de résultats de l'analyse comparative des modèles de régression de Cox pour BADRFI, NP et C-index-OP sur les performances prédictives de la survie globale à 5 ans pour la cohorte de première ligne R/T du CHU de Bordeaux.	166
5.18	Distribution des résultats de performances des 22 cohortes de l'HEGP (à gauche) et des 22 cohortes du CHU de Bordeaux (à droite). Les barres verticales quantifient le nombre de cohortes alignées sur des combinaisons spécifiques de résultats de performance. Ces combinaisons de performances sont illustrées dans la matrice codée par couleur en-dessous. La couleur des points indique la performance relative : vert pour les gains, rouge pour les pertes et noir pour les résultats équilibrés. Les barres horizontales montrent les nombres agrégés de cohortes connaissant des gains, des pertes ou des résultats équilibrés.	167

5.19	Trois évaluations des performances relatives de NP comparées à BADRDI sur les 22 cohortes de l'HEGP (à gauche) et les 22 cohortes du CHU de Bordeaux (à droite). Pour chaque cohorte, les barres à droite indiquent un gain, les barres à gauche indiquent une perte, et l'absence de barre indique un résultat équilibré dans la métrique d'évaluation correspondante.	169
5.20	Gains des scores de prédiction de la survie globale à 5 ans des régressions de Cox (C-index) et logistique (AUC-ROC)	171
5.21	Gains de prédiction en C-index obtenus avec les modèles ProtoDrift par rapport au modèle BADRDI selon les valeurs de α (i.e., $\frac{\omega_t}{\omega_t + \omega_d}$) et β (i.e., $\frac{\omega_{\text{inter}}}{\omega_{\text{inter}} + \omega_{\text{intra}}}$) avec la cohorte respiratoire et thoracique en première ligne à l'HEGP (à gauche, figure 5.21a) et au CHU de Bordeaux (à droite, figure 5.21b).	172
5.22	Graphiques en facettes illustrant les régions des <i>heatmaps</i> avec un gain maximal pour les cohortes ayant obtenu un gain de prédiction supérieur à 1%. À gauche, les graphiques présentent les régions avec un gain maximal en AUC-ROC. À droite, les graphiques présentent les régions avec un gain maximal en C-index. Chaque rangée de graphiques représente une ligne de traitement et un temps de survie spécifiques. Les couleurs représentent les différentes localisations de cancer.	176
5.23	Comparaison entre notre méthode d'évaluation de ProtoDrift et la méthode de validation externe recommandée dans la littérature. Figure adaptée de l'article de SURESH, SEVERN et GHOSH [195].	179

LISTE DES TABLEAUX

2.1	Représentations de connaissances bio-médicales mentionnées dans cette thèse.	42
2.2	Algèbre d'Allen. X et Y décrivent deux intervalles temporels, délimités par leurs instants de début X_{start} , Y_{start} et leurs instants de fins X_{end} , Y_{end}	53
3.1	Format des données de survie avec covariables pour les patients 1, 2 et 3 représentés dans la figure 3.4. Les covariables peuvent inclure des caractéristiques cliniques comme l'âge, le sexe, et d'autres informations spécifiques aux patients. Par exemple, dans l'étude de MENG et al. [131], les covariables incluent l'âge, le sexe, la présence de métastases, la chirurgie de la tumeur primitive, et le nombre de sites métastatiques. Dans l'étude de BREADNER et al. [24], les covariables incluent l'âge, le sexe, la dose-intensité relative (RDI), le stade tumoral et le stade nodal.	72
4.1	Triplets possibles dans ChemoOnto. Les entités de ChemoOnto sont surlignées en rouge, celles de la Time Ontology en bleu et celles de Romedi en orange. Les littéraux sont écrits en italique.	99
4.2	Suite des triplets possibles dans ChemoOnto. Les entités de ChemoOnto sont surlignées en rouge, celles de la Time Ontology en bleu et celles de Romedi en orange. Les littéraux sont écrits en italique.	100
4.3	Comparaison des classes et propriétés de ChemoOnto avec des ontologies existantes	104
4.4	Répartition des instances des classes de ChemoOnto à l'HEGP et au CHU de Bordeaux	119
4.5	Répartition des instances des classes d'OntoTox extraites de trois sources différentes : les questionnaires cliniques, les textes libres et les tables semi-structurées. Cette table montre le nombre total de toxicités et de grades détectés dans chaque source. On rappelle que "GradeNull" fait référence à un grade dont on n'a pas pu détecter le numéro, tandis que "Grade0" fait référence à l'absence explicite de la toxicité.	121
4.6	Classes d'OntoTox et nombres d'instances	123
4.7	Classes de ChemoOnto et nombres d'instances	123
4.8	Instances de la propriété d'objet time :inside	124
5.1	Modèles de régression logistique pour BADRDI, NP et AUC-ROC-OP . . .	152
5.2	Modèles de régression de Cox pour BADRDI, NP et C-index-OP	152
5.3	Correspondance des noms de localisations de cancers/maladies entre HEGP et CHU de Bordeaux. Les appellation conservées pour notre étude sont celles de l'HEGP.	162
5.5	Prédiction des ajustements du modèle de Cox et détails gain/perte du pouvoir discriminant. Table de l'HEGP.	191
5.7	Prédiction des ajustements du modèle de Cox et détails gain/perte du pouvoir discriminant. Table du CHU de Boredaux.	194
5.9	HEGP : Détails du pouvoir discriminant (1/3)	195
5.11	HEGP : Détails du pouvoir discriminant (2/3)	196

5.13	HEGP : Détails du pouvoir discriminant (3/3)	197
5.15	Bordeaux : Détails du pouvoir discriminant (1/3)	198
5.17	Bordeaux : Détails du pouvoir discriminant (2/3)	199
5.19	Bordeaux : Détails du pouvoir discriminant (3/3)	200

REMERCIEMENTS

Je tiens avant tout à exprimer ma profonde gratitude envers mes directeurs de thèse, Adrien Coulet et Bastien Rance, pour leur confiance et leur soutien constants au fil de ces quatre années. Leur accompagnement, leurs conseils avisés et leurs encouragements m'ont permis de mener à bien cette aventure scientifique.

Je remercie vivement Brigitte Séroussi et Fleur Mougin d'avoir accepté d'être les rapportrices de ce manuscrit, ainsi que pour l'intérêt qu'elles ont porté à mon travail. J'adresse également mes plus sincères remerciements à Sandra Bringay, Marie-Anne Loriot, Maud Toulmonde et Vianney Jouhet pour avoir accepté de faire partie du jury de ma soutenance.

Ma reconnaissance va également aux trois équipes avec lesquelles j'ai eu la chance de collaborer durant cette thèse :

Merci à tous mes collègues de l'HEGP, notamment Hector, Éric, David, Valentin, Ivan, Pierre, Joséphine, William et Maxime, pour les déjeuners récréatifs qui ont allégé les journées de travail.

Je tiens à exprimer ma gratitude aux membres de l'équipe HeKA, en particulier Antoine, Diana, Jong Ho, Lillian, Safa, Fabien, Camila, Solange, Emma, Jean, Ghislain, Moreno, Anne-Sophie et Sarah. Merci pour toutes ces discussions partagées autour d'un café ou d'une bière sur le *rooftop*, quand nous n'avions pas encore la possibilité de les avoir autour d'un repas à la cantine.

À Louis et Linus : tout soutien.

Je suis également très reconnaissante envers Romain Griffier et Corine Barat ainsi que toute l'équipe UIAM de Bordeaux pour la semaine de collaboration enrichissante, dans une ambiance bienveillante.

Un immense merci à Juliette Murris et Rosy Tsopra, véritables sources d'inspiration et de soutien tout au long de cette thèse.

Au-delà du travail, je suis extrêmement chanceuse d'être entourée de personnes formidables.

Merci à l'association MLJ Danse, la meilleure association de modern jazz de Paris, pour les moments de joie et d'expression artistique qu'elle m'offre.

Je remercie chaleureusement ma famille, et en particulier mes parents et ma sœur, pour leur passion et leur amour, qui sont pour moi une force constante.

J'ai aussi une grande famille de cœur. Merci aux amis de longue date, de la cohabitation à La Plata aux quinze voire vingt ans d'amitié partagée avec la MDTeam. Notre amitié est ma plus grande fierté. Un clin d'œil spécial à Madeleine, la meilleure prof de SVT, et mes trois inséparables Mathilde, Audrey et Rachel, sandales de ma vie.

Et enfin, à toi, Victor, pour tout ce que tu représentes.

CHAPITRE 0

RÉSUMÉ

La chimiothérapie est le traitement prédominant pour soigner les cancers. C'est un traitement complexe au cours duquel des molécules cytotoxiques sont administrées au patient pour cibler les cellules cancéreuses et réduire la taille des tumeurs. Cependant, la réponse des patients à ces traitements varie considérablement, et l'administration de ces molécules entraîne souvent des toxicités de gravité variable. Mieux comprendre et gérer ces réponses individuelles est un sujet de recherche majeur en oncologie. Ainsi, extraire les informations sur les toxicités et évaluer l'efficacité du traitement sont des étapes essentielles dans les études qui visent à améliorer et personnaliser le traitement des cancers.

L'objectif de cette thèse est d'exploiter les données hospitalières afin d'extraire, représenter et prédire les réponses aux chimiothérapies.

Les informations individuelles sur les chimiothérapies suivies et leurs réponses sont disponibles dans les entrepôts de données hospitaliers, mais leur extraction présente plusieurs obstacles. Celles-ci sont irrégulièrement réparties et structurées dans des sources hétérogènes, telles que des questionnaires, les données de suivi des chimiothérapies et des comptes rendus. De plus, les chimiothérapies suivent des protocoles complexes incluant les notions de cycles, lignes de traitement, de changements de doses et de temps d'administration qui ne sont pas facilement manipulables avec les entrepôts de données classiques.

Pour répondre à ces défis, nous avons développé deux ontologies : ChemoOnto et OntoTox. ChemoOnto est une ontologie dédiée à la représentation du cours des chimiothérapies, en intégrant à la fois les protocoles théoriques et les lignes de traitement suivies par les patients. Elle réutilise des ontologies de référence pour représenter la temporalité et les médicaments. OntoTox, quant à elle, représente les toxicités induites par les chimiothérapies et leurs sévérités, extraites à partir de diverses parties d'un entrepôt de données hospitalier. OntoTox normalise ces toxicités en utilisant des thésaurus médicaux standardisés. Les deux ontologies ont été peuplées avec des données réelles, constituant ainsi deux bases de connaissances qui facilitent l'analyse de données sur les chimiothérapies et les réponses à ces traitements. Ces ressources ont été implémentées avec les outils du web sémantique qui permettent de surmonter certaines limites des entrepôts de données traditionnels en offrant une représentation standard et extensible des connaissances biomédicales.

En parallèle, nous avons développé ProtoDrift, une métrique pour mesurer l'adhésion aux chimiothérapies. ProtoDrift évalue l'écart entre le protocole de chimiothérapie théoriquement suivi et le traitement effectivement suivi par un patient, en prenant en compte les décalages de jours d'administration et les réductions de doses. Cette métrique utilise une moyenne pondérée des dissimilarités à plusieurs niveaux (administration, molécule, cycle, ligne de traitement). L'optimisation des poids de ProtoDrift par des régressions logistiques ou de Cox montre une amélioration significative des prédictions de survie par rapport à la méthode classique de mesure de l'adhésion aux chimiothérapies, la dose intensité relative. Ces résultats ont été reproduits dans deux hôpitaux différents, confirmant la robustesse et la pertinence de ProtoDrift pour évaluer l'adhésion aux chimiothérapies. Cette reproductibilité a notamment été facilitée par ChemoOnto que nous avons utilisé comme représentation standard pour calculer ProtoDrift.

Les ontologies ChemoOnto et OntoTox, ainsi que la métrique ProtoDrift, offrent des perspectives prometteuses pour la recherche en oncologie. Elles facilitent la considération de données médicales complexes et permettent une analyse plus fine des réponses aux chimiothérapies. De plus, ces ressources ont été développées en suivant les principes et bonnes pratiques de la science ouverte.

Mots-clés : chimiothérapie, données de vie réelle, toxicités, adhésion, fouille de données, graphes de connaissances, survie

ABSTRACT

Chemotherapy is the primary treatment for cancer, involving the administration of cytotoxic drugs to patients to target cancer cells and reduce tumor size. However, patient response to this treatment varies widely, and its administration often results in toxicities of various levels of severity. Better understanding and managing these individual responses is a major focus in oncology research. Extracting toxicities and assessing treatment efficacy are crucial for studies that aim at improving and personalizing cancer treatments.

The goal of this thesis is to leverage hospital data to extract, represent, and predict chemotherapy responses.

Information about chemotherapies and their responses is available in hospital data warehouses, but its use faces several challenges. This information is irregularly distributed and structured across heterogeneous sources such as questionnaires, chemotherapy follow-up data, and medical reports, making its extraction difficult. Additionally, chemotherapies involve complex concepts such as treatment cycles, lines, theoretical protocols, and toxicity grades, which are not easily handled by classical data warehouses.

To address these challenges, we developed two ontologies : ChemoOnto and OntoTox. ChemoOnto is designed to represent the course of chemotherapy treatments, integrating both theoretical protocols and actual patient treatment lines. It employs reference ontologies to represent temporal aspects and drugs. OntoTox, on the other hand, focuses on the chemotherapy-induced toxicities and their severity, extracted from various parts of a hospital data warehouse. OntoTox normalizes these toxicities using standard medical thesauri. Both ontologies were populated with real-world data to form knowledge bases that facilitate the analysis of chemotherapy data and outcomes. Additionally, they were implemented using Semantic Web tools to overcome the limitations of classical data warehouses by providing a standard and extensible representation of biomedical knowledge.

Meanwhile, we developed ProtoDrift, a metric to measure the adherence to chemotherapy protocols. ProtoDrift evaluates deviations between the theoretical chemotherapy protocol supposed to be followed and the actual one, by considering administration day shifts and dose reductions. This metric uses a weighted average of dissimilarities at several levels (administration, molecule, cycle, treatment line). Optimizing ProtoDrift weights using logistic regression or Cox methods has shown significant improvements in survival predictions compared to the traditional measure of chemotherapy adherence, namely the relative dose intensity. These results were reproduced in two different hospitals, confirming the robustness and relevance of ProtoDrift in evaluating chemotherapy adherence. In addition, we relied on ChemoOnto to serve as a standard representation to compute ProtoDrift in the two settings.

The ChemoOnto and OntoTox ontologies, along with the ProtoDrift metric, offer promising perspectives for oncology research. They facilitate the consideration of complex medical data and enable a fine-grained understanding of chemotherapy responses. Lastly, these resources have been developed using Open Science principles and good practices.

Keywords : chemotherapy, real-world data, toxicities, adherence, data mining, knowledge graphs, survival

GLOSSAIRE

ADN Acide désoxyribonucléique.

ADRDI *All Drugs Relative-Dose Intensity* Dose intensité-relative pour tous les médicaments d'un schéma thérapeutique (*All Drugs Relative Dose Intensity*).

AP-HP Assistance Publique Hopitaux de Paris.

AUC-ROC Aire sous la courbe de ROC. Score de prédiction de la régression logistique.

AUC-ROC-OP Méthode de mesure de l'adhésion aux chimiothérapies utilisant ProtoDrift Optimisé par aire sous la courbe de ROC (régression logistique).

BADRDI La méthode de la ligne de base pour mesurer l'adhésion aux chimiothérapies utilisant la dose-intensité relative pour tous les médicaments d'un schéma thérapeutique (*Baseline All Drugs Relative-Dose Intensity*).

C-index-OP Méthode de mesure de l'adhésion aux chimiothérapies utilisant ProtoDrift Optimisé par C-Index (régression de Cox).

Chimio® Logiciel de suivi des chimiothérapies utilisé dans de nombreux hôpitaux dont l'HEGP et le CHU de Bordeaux.

CHU Centre Hôpitalo-Universitaire.

CTCAE *Common Terminology Criteria for Adverse Events*. Classification permettant de écrire et quantifier la sévérité des toxicités chimio-induites.

DPI Dossiers Patients Informatisés.

ETL Processus d'extraction, transformation et chargement (pour *Extraction Transformation Loading*).

GCN *Graph Convolutional Networks*.

HALP Score hémoglobine, albumine, lymphocytes, plaquettes.

HAS Haute Autorité de Santé.

HEGP Hôpital Européen Georges Pompidou.

i2b2 *Informatics for Integrating Biology and the Bedside*.

INCa Institut National du Cancer.

INR *International Normalised Ratio*.

KM Kaplan-Meier.

LASSO *Least Absolute Shrinkage and Selection Operator*.

NP ProtoDrift Naïf. Méthode de mesure de l'adhésion aux chimiothérapies utilisant ProtoDrift sans optimiser les poids associés aux différentes dissimilarités..

NSCLC Cancer du poumon non à petites cellules (pour *non-small cell lung cancer*).

OMS Organisation Mondiale de la Santé.

PMSI Programme de Médicalisation des Systèmes d'Information.

PS-OP Méthode de mesure de l'adhésion aux chimiothérapies utilisant ProtoDrift Optimisé par un Score de Prédition.

RCP Réunion de Concertation Pluridisciplinaire.

RDI Dose-Intensité Relative.

REN Reconnaissance d'Entités Nommées.

SG Survie Globale. La survie globale mesure le temps écoulé entre le début du suivi et le décès..

SIH Système d'Information Hospitalier.

SNOMED CT *Systematized Nomenclature of Medicine Clinical Terms* (ontologie).

SSP Survie Sans Progression. La survie sans progression mesure le temps écoulé entre le début du traitement et la progression de la maladie ou le décès dû à toute cause.

SSR Survie sans récidive.

SWRL *Semantic Web Rule Language*.

TAL Traitement Automatique des Langues.

TNM Classification fournissant une description détaillée de l'extension du cancer à partir des trois phénomènes Tumeur, Nœud et Métastase.

INTRODUCTION

Le cancer est une maladie provoquée par la transformation de cellules qui deviennent anormales et se multiplient de façon anarchique. Elles forment alors une masse que l'on appelle tumeur maligne. À des stades avancés de la maladie, ces cellules dérégées migrent par les vaisseaux sanguins et lymphatiques et forment des tumeurs secondaires, les métastases. Les traitements contre les cancers visent à limiter ces progressions [140].

Malgré les avancées récentes sur les immunothérapies, les chimiothérapies, en combinaison avec la chirurgie et la radiothérapie, restent le traitement le plus courant pour soigner les cancers [28]. Elles consistent à éliminer les cellules cancéreuses en administrant des médicaments cytotoxiques. Ces médicaments ciblent les cellules qui se divisent rapidement, un trait caractéristique des cellules cancéreuses. Cependant, ils affectent également les cellules saines, ce qui entraîne la survenue d'effets indésirables, aussi appelées toxicités.

Dans les communautés médicales, l'utilisation du terme "maladie chronique" pour qualifier le cancer fait débat [159]. L'OMS (Organisation Mondiale de la Santé) définit une maladie chronique « comme une affection de longue durée qui, en règle générale, évolue lentement. » Une affection de longue durée est quant à elle définie comme une maladie grave ou chronique nécessitant un traitement prolongé. Ces définitions sont subjectives et sujettes à interprétation, mais dans la population générale, le cancer est considéré comme une maladie grave et la chimiothérapie comme longue, éreintante et partiellement nocive [126]. De nombreuses études et actions sont menées pour améliorer la qualité de vie des patients sous chimiothérapies [104], car leurs répercussions indésirables sont souvent inévitables pour lutter contre le cancer.

Les réponses aux chimiothérapies se réfèrent aux résultats cliniques observés après l'administration de médicaments. Elles incluent d'une part l'efficacité du traitement, c'est-à-dire la limitation de la progression du cancer, et d'autre part la survenue de toxicités.

Ces réponses varient d'un patient à l'autre. Elles dépendent de nombreux facteurs individuels tels que l'âge, les comorbidités ou le profil génétique [147, 191, 167, 56]. Deux patients aux profils différents ne réagissent pas de la même façon à l'administration de la même chimiothérapie.

En fonction de leurs connaissances et de leurs expériences, les oncologues prévoient les réponses des patients. Ils adaptent les traitements pour limiter la survenue de toxicités ou pour intensifier la réduction de la masse tumorale [74]. L'adaptation des traitements aux caractéristiques individuelles des patients est au cœur de la médecine de Personnalisation, Précision, Prévention, dite "médecine des 3P" (*Predictive, Preventive & Personalised Medicine - PPPM*) [72]. Or, beaucoup de facteurs influençant les réponses des patients sont méconnus. Ainsi, la personnalisation des traitements de chimiothérapie est un thème de recherche actif en oncologie [127, 69].

En parallèle, l'informatisation des systèmes d'information a contribué au développement des entrepôts de données hospitaliers. Ceux-ci fournissent une source d'informations sur les chimiothérapies, leurs réponses et les profils des patients. Ils stockent une vaste

quantité de données provenant de diverses sources, comme les comptes rendus de consultation, les résultats de laboratoires, les questionnaires, et les données de suivi des chimiothérapies.

Dans ce contexte, les entrepôts de données hospitaliers constituent un matériel précieux et unique pour étudier les réponses aux chimiothérapies. Cependant, l'exploitation des données hospitalières pose plusieurs défis en raison de leur volume, de leur complexité et de leur diversité. Les informations sont dispersées dans plusieurs sources de données aux formats variés. Ceci rend leur réutilisation pour la recherche complexe.

Avant de réutiliser les données hospitalières pour la recherche, il faut surmonter plusieurs obstacles. Il faut préparer les données pour en extraire des connaissances. Cette préparation inclut premièrement l'identification d'informations d'intérêt. Deuxièmement, il faut normaliser ces informations à l'aide de représentations de connaissances formelles de manière à ce qu'elles soient à la fois lisibles par une machine et un humain. Cette représentation permet de comparer plus facilement les connaissances extraites. Il est alors possible de tester des hypothèses avec des modèles prédictifs, et de contribuer à l'analyse des réponses aux chimiothérapies.

Cette thèse a pour objectif de développer des outils d'extraction de connaissances et de prédiction des réponses aux chimiothérapies à partir des données hospitalières.

Nous répondons à ces défis avec deux contributions principales :

- nous proposons deux ontologies, **ChemoOnto** et **OntoTox** pour représenter les connaissances sur les chimiothérapies et leurs réponses extraites depuis les entrepôts de données hospitaliers.
- nous proposons **ProtoDrift**, une nouvelle mesure de l'adhésion aux chimiothérapies.

ChemoOnto [170] est une ontologie dédiée à la représentation des connaissances extraites sur le cours des chimiothérapies. Cette ontologie permet de représenter à la fois les protocoles théoriques de chimiothérapie et les traitements effectivement suivis par les patients, facilitant ainsi leur comparaison. ChemoOnto est reliée à des ontologies de référence pour représenter la temporalité et les médicaments. Elle a été instanciée avec les données de deux hôpitaux, l'Hôpital Européen Georges Pompidou (HEGP) et le Centre Hospitalo-Universitaire de Bordeaux (CHU de Bordeaux), démontrant ainsi sa portabilité. La représentation des protocoles théoriques de ChemoOnto a été intégrée au sein d'un graphe de connaissance, **ChemoKG** [105] disponible en ligne sur <http://chemokg.paris.inria.fr/>. L'utilisation d'algorithmes de plongement de graphes sur ChemoKG a permis de clusteriser les protocoles théoriques de chimiothérapie de manière cohérente.

De son côté, **OntoTox** [169] est une ontologie dédiée à la représentation des toxicités et de leurs sévérités, extraites à partir de différentes sources de données hospitalières. Elle est également reliée à des terminologies de référence permettant de normaliser les toxicités extraites et d'assurer leur traçabilité. Elle a été instanciée avec les données de l'entrepôt de l'HEGP, sur une cohorte de patients atteints de cancer du poumon, ce qui a démontré sa capacité à intégrer et structurer les informations sur les toxicités provenant de différentes

sources.

Les deux ontologies instanciées forment des bases de connaissances qui facilitent l'analyse de données sur les chimiothérapies et leurs réponses. Enfin, elles ont été reliées entre elles par des relations temporelles.

ProtoDrift est une nouvelle mesure de l'adhésion aux chimiothérapies, qui surmonte les limitations de la mesure classique, la dose intensité relative (*relative dose intensity*, RDI), notamment en prenant mieux en compte les aspects calendaires des traitements. ProtoDrift quantifie les écarts par rapport aux protocoles de traitement prévus en sommant des dissimilarités pondérées à différents niveaux du traitement de chimiothérapie. En optimisant les poids associés à ces dissimilarités, ProtoDrift permet de capturer les impacts des ajustements de traitement sur la survie des patients. Cette méthode a été appliquée aux données des patients de l'HEGP et du CHU de Bordeaux pour analyser les écarts aux protocoles théoriques. Elle a généré des résultats pan-cancer (<https://files.inria.fr/protodrift-surv/>). On observe dans ces résultats, différents types de déviations aux chimiothérapies selon les localisations de cancer, ce qui souligne l'importance d'une adaptation des poids de ProtoDrift à chaque contexte spécifique.

Le présent manuscrit se compose de cinq chapitres répartis en quatre parties.

I. Une première partie "Contexte" est constituée du premier chapitre :

1. Le premier chapitre contextualise le sujet de la thèse. Il présente d'une part les chimiothérapies, et d'autre part les entrepôts de données hospitaliers.

II. Une deuxième partie "État de l'art" est composée de deux chapitres :

- 2 Le deuxième chapitre porte sur la représentation des connaissances. Il présente les ontologies et montre leur intérêt pour représenter les connaissances biomédicales.
- 3 Le troisième chapitre porte sur l'extraction des connaissances à partir de diverses sources, et la prédiction d'événements avec des modèles de survie.

III. Une troisième partie "Contributions" est composée de deux chapitres :

- 4 Le quatrième chapitre présente notre première contribution : le développement des ontologies ChemoOnto et OntoTox pour représenter les chimiothérapies et leurs réponses extraites à partir de diverses sources de données.
- 5 Le cinquième chapitre présente notre deuxième contribution : la conception d'une mesure d'adhésion aux chimiothérapies, ProtoDrift.

IV. Enfin une dernière partie conclut le manuscrit.

CHAPITRE 0

Première partie

Contexte

CHAPITRE 1

LES CHIMIOTHÉRAPIES À L'HEURE DES ENTREPÔTS DE DONNÉES

1.1	Le principe des chimiothérapies	16
1.1.1	Le cancer et sa progression	16
1.1.2	La place des chimiothérapies dans le traitement du cancer	17
1.1.3	L'action des molécules anticancéreuses des chimiothérapies	18
1.1.4	Les schémas thérapeutiques de chimiothérapies	19
1.2	Les chimiothérapies dans la pratique	20
1.2.1	Les réponses patients-dépendantes des chimiothérapies	20
1.2.2	Le déroulement des chimiothérapies	22
1.2.3	Le suivi des réponses aux chimiothérapies	23
1.2.4	L'adaptation des chimiothérapies aux réponses des patients	24
1.2.5	L'adhésion aux chimiothérapies	24
1.2.6	Les adaptations praticien-dépendantes aux protocoles	25
1.3	Évaluation des réponses aux chimiothérapies dans les études cliniques	26
1.3.1	Évaluation des schémas thérapeutiques	26
1.3.2	Évaluation de l'adhésion aux schémas thérapeutiques	26
1.3.3	La nécessité d'un consensus sur les schémas thérapeutiques et le potentiel des DPI	27
1.4	Les entrepôts de données de santé	28
1.4.1	Les données de santé et les Dossiers Patients Informatisés	28
1.4.2	La provenance et l'hétérogénéité des données des DPI	28
1.4.3	L'alimentation des entrepôts de données hospitaliers	29
1.4.4	Le schéma en étoile d'i2b2	29
1.4.5	Les entrepôts de données de l'HEGP et du CHU de Bordeaux	29
1.4.6	Les enregistrements sur les chimiothérapies	30
1.4.7	Les enregistrements de réponses aux chimiothérapies	30
1.4.8	Les données de décès de l'Insee	32
1.5	Les défis à relever pour réutiliser les DPI à des fins de recherche	32
1.5.1	Données dédiées à la recherche vs. données de vie réelles réutilisées pour la recherche	32

1.5.2	La qualité des DPI et leur validité	32
1.5.3	La nécessité d'extraire, de représenter et d'intégrer les informations des DPI	32
1.5.4	Différence de conclusion et implications des résultats obtenus avec des DPI	33
1.6	Bilan : Les défis de la réutilisation des DPI pour étudier les réponses aux chimiothérapies	33



Ce chapitre présente de façon simple deux domaines : les chimiothérapies (sections 1.1 à 1.3) et les données médicales (sections 1.4 à 1.5). Il présente la complexité des chimiothérapies que l'on cherche à représenter (sections 1.1 et 1.2), l'évaluation des réponses dans la pratique (sections 1.2.3 à 1.2.6) et la recherche clinique (section 1.3), sur lesquelles on s'appuie pour développer des méthodes d'extraction et de prédiction, et notre matériel d'étude : les Dossiers Patients Informatisés (DPI) (section 1.4). Il se termine par les défis à relever pour réutiliser des données de patients à des fins de recherche (section 1.5).

Les sections 1.1 et 1.2 s'appuient sur des informations présentes sur les sites de l'institut national du cancer (l'INCa) [97], de l'*American Cancer Society* [5], du collège national de pharmacologie médicale [40] et sur la thèse de Clémence Poirot [161].

1.1 Le principe des chimiothérapies

1.1.1 Le cancer et sa progression

Le cancer se caractérise par la transformation de cellules normales en cellules anormales qui se multiplient de manière incontrôlée. Cette prolifération anarchique entraîne la formation de masses appelées tumeurs malignes. Les connaissances de la biologie du cancer concernent essentiellement la découverte de mutations génétiques qui altèrent le comportement cellulaire normal. Les gènes responsables de la régulation de la croissance cellulaire, de la réparation de l'ADN et de la mort cellulaire programmée (apoptose) subissent des altérations. Ces mutations peuvent être héritées ou acquises au cours de la vie en raison de facteurs environnementaux, tels que le tabagisme, les radiations, les infections virales, etc. Les cancers sont classés en trois grands groupes, en fonction de leur origine tissulaire.

- Les carcinomes, qui représentent la majorité des cancers, se développent à partir des cellules épithéliales. Parmi eux, on trouve les cancers du sein, de la prostate, du poumon et du colon.
- Les sarcomes, bien que moins fréquents, prennent naissance dans les tissus conjonctifs ou de soutien, comme les os, les muscles et les cartilages.
- Les cancers hématologiques, comme les leucémies, les lymphomes et les myélomes, affectent les cellules sanguines et les tissus formateurs de sang. Les lymphomes, une sous-catégorie des cancers hématologiques, incluent les lymphomes de Hodgkin et non-Hodgkiniens.

La progression d'un cancer est décrite par son stade, qui reflète l'étendue de la maladie. Le stade d'un cancer permet de quantifier la progression et l'étendue de la maladie. Au

stade I, la tumeur primitive est généralement localisée et de petite taille. Au stade II, la tumeur est plus grande et les cellules cancéreuses peuvent commencer à envahir les ganglions lymphatiques proches de la tumeur sous forme de tissus anormaux, créant des nodules, des masses ou des ulcérations¹, qui sont appelés “lésions”. Au stade III, le cancer montre une extension locale plus importante, atteignant davantage de ganglions lymphatiques avec des lésions plus nombreuses et étendues. Le stade IV, le plus avancé, indique la présence de métastases², où les cellules cancéreuses se sont propagées à d’autres parties du corps via les systèmes sanguin et lymphatique, formant des tumeurs secondaires.

1.1.2 La place des chimiothérapies dans le traitement du cancer

Le traitement du cancer regroupe l’ensemble des soins, utilisés seuls ou en combinaison, pour confirmer le diagnostic, réduire ou éliminer la tumeur primitive et les métastases, limiter le risque de récidive, soulager les symptômes et prolonger la vie des patients. Le choix des traitements dépend des objectifs médicaux, du type et du stade du cancer et des caractéristiques des patients. Ces traitements se caractérisent essentiellement par les moyens employés pour atteindre ces objectifs.

- La chimiothérapie, au sens propre, s’appuie sur l’administration de molécules cytotoxiques pour détruire les cellules tumorales en endommageant leur ADN. Depuis une dizaine d’années, les chimiothérapies incluent parfois des thérapies ciblées. Celles-ci consistent en l’administration de molécules qui interfèrent avec des cibles moléculaires essentielles au développement et à la progression des tumeurs. Les traitements qui combinent une chimiothérapie et une thérapie ciblée sont souvent appelés chimiothérapie dans un sens élargi.
- La chirurgie permet de confirmer le diagnostic malin d’une tumeur ou de réduire partiellement ou totalement la tumeur par ablation.
- La radiothérapie emploie des rayonnements ionisants pour détruire les cellules cancéreuses.
- L’hormonothérapie inhibe les hormones qui favorisent la croissance tumorale.
- L’immunothérapie utilise des anticorps pour activer le système immunitaire contre les cellules cancéreuses.

Les chimiothérapies peuvent être administrées en complément d’une chirurgie. La chimiothérapie pré-opératoire, dite néoadjuvant a pour but de réduire la taille de la tumeur à enlever. La chimiothérapie post-opératoire, dite chimiothérapie adjuvante, a pour objectif de diminuer les risques de récidives. La chimiothérapie peut aussi parfois être associée à une radiothérapie, l’objectif étant de diminuer la zone à irradier.

Selon l’INCa, en France, l’activité de chimiothérapie représente 40.8% des prises en charges des cancers dans les établissements publics habilités à faire de la chimiothérapie en 2017 [28]. Avec les avancées de l’immunothérapie, son recours tend à diminuer mais elle reste un des traitements les plus fréquents pour traiter le cancer. D’après le panorama des cancers de l’INCa [99], sur 433 136 nouveaux cas de cancers en 2023, 363 160 soit près

1. Une ulcération est une lésion locale et superficielle

2. Tumeur formée à partir de cellules cancéreuses qui se sont détachées de la tumeur primitive et qui ont migré par les vaisseaux lymphatiques ou les vaisseaux sanguins dans une autre partie du corps où elles se sont installées.

de 84%, sont soignés par une chimiothérapie (seule ou combinée, le plus souvent à une chirurgie).

1.1.3 L'action des molécules anticancéreuses des chimiothérapies

Les molécules cytotoxiques administrées dans les chimiothérapies endommagent l'ADN des cellules en division lors des mitoses, et provoquent leur mort cellulaire immédiate ou programmée, c'est-à-dire leur apoptose. Elles se divisent en deux catégories suivant leur interaction directe ou indirecte avec l'ADN. Ces deux catégories sont elles-mêmes divisées en familles de molécules, réparties selon leur mode d'action pour léser l'ADN. Les molécules qui interagissent directement sur l'ADN se divisent en trois familles [161] :

- les agents alkylant qui troublent la réPLICATION et la transcription de l'ADN en formant des liaisons covalentes (par exemple Oxaliplatine, Carboplatine et Cisplatine) ;
- les agents intercalants qui s'insèrent entre les brins d'ADN et empêchent aussi sa réPLICATION et sa transcription (par exemple Mitoxantrone, Doxorubicine et Dactinomycine) ;
- l'agent scindant, appelé Bléomonycine, qui lui “casse” l'ADN.

Les molécules qui altèrent indirectement l'ADN se divisent en quatre familles :

- les antimétabolites qui perturbent la synthèse des bases de l'ADN ;
- les inhibiteurs d'enzymes essentielles à la transcription et réPLICATIONS de l'ADN ;
- les antimitotiques qui inhibent des enzymes essentielles à la mitose ;
- d'autres molécules diverses n'appartenant pas à une famille spécifique inhibent ou activent des enzymes spécifiques, avec toujours pour conséquence des lésions de l'ADN (par exemple Asparaginase, Hydroxycarbamide et Bortezomib).

La chimiothérapie inclut parfois une thérapie ciblée. Les molécules administrées dans les thérapies ciblées se distinguent par leur capacité à intervenir spécifiquement sur des composantes biologiques essentielles à la survie et la prolifération des cellules cancéreuses. Ces thérapies sont dites ciblées, car les molécules associées visent des cibles moléculaires précises, souvent impliquées dans les voies de signalisation cellulaire. Ce mode d'action est distinct de celui des molécules cytotoxiques qui agissent de manière plus générale sur les cellules en division. Dans une thérapie ciblée, les molécules administrées se divisent en deux catégories :

- les anticorps monoclonaux qui se lient à des antigènes présents en grande quantité à la surface des cellules tumorales (par exemple Trastuzumab, béravacizumab et Rituximab)
- les inhibiteurs de la tyrosine kynase, une enzyme présente en plus grande quantité dans les cellules tumorales et qui favorisent leur croissance (par exemple Erlotinib, Lapatinib et Ruxolitinib)

Cette classification permet non seulement de comprendre les mécanismes d'action des traitements mais aussi de prévoir les groupes d'effets indésirables associés, tels que les cardiotoxicités pour les agents alkylants [157] ou l'hypertension pour les thérapies ciblées [117]. On montrera dans le chapitre suivant que les inter-connections de classifications dans le domaine médical permettent d'enrichir les graphes de connaissances.

1.1.4 Les schémas thérapeutiques de chimiothérapies

Les cellules tumorales, caractérisées par leur division rapide, sont particulièrement “chimiosensibles”, c'est à dire sensible aux anticancéreux, et donc préférentiellement éliminées par les molécules anticancéreuses. Cependant, les molécules anticancéreuses ne ciblent pas exclusivement ces cellules malignes ; elles affectent également les cellules saines, engendrant ainsi divers effets indésirables. La chimiothérapie repose sur un équilibre nuancé entre la réduction de la tumeur et la minimisation des effets indésirables. Pour trouver cet équilibre, les chimiothérapies impliquent souvent l'administration de combinaison de deux à trois molécules anticancéreuses distinctes [214]. Ces molécules sont administrées de manière cyclique, alignée avec leur demi-vie, *i.e.* leur durée d'activité dans l'organisme. Un schéma thérapeutique décrit un cycle, c'est-à-dire les molécules, leur modalité d'administration, leur temporalité d'administration. Ainsi la définition d'un cycle comprend pour chaque molécule :

- la dose administrée, généralement exprimée en mg/m²
- le ou les jours d'administration dans le cycle,
- la voie d'administration, souvent intraveineuse, intramusculaire, ou orale,
- et, pour les administrations intraveineuses, le mode d'infusion (rapide, dite “bolus” ou perfusion lente).

Ces paramètres sont établis lors de la phase 1 des essais cliniques, grâce à l'utilisation de modèles pharmacocinétiques et pharmacodynamiques (PK/PD). Le but de ces paramètres est d'atteindre des concentrations sanguines qui maximisent l'efficacité thérapeutique tout en minimisant les risques d'effets indésirables. Historiquement, le modèle de la dose maximale tolérée [8] a été le premier à être utilisé, et reste le modèle classiquement utilisé aujourd'hui. Il consiste à déterminer la dose maximale supportable avant l'apparition d'effets indésirables jugés trop graves, dans un intervalle de temps donné. Depuis, d'autres modèles tels que la dose biologique effective [175] et la dose biologique optimale [67] sont également employés. Ces derniers cherchent à déterminer la dose minimale nécessaire pour obtenir l'effet biologique souhaité avec le moins de toxicité possible, en utilisant des approches statistiques plus avancées. Des modèles innovants sont également à l'étude, comme ceux basés sur la chimiothérapie métronome [17], visant à optimiser le traitement des cancers métastatiques par des administrations fréquentes de faibles doses de molécules anticancéreuses, afin de limiter la progression des métastases et les effets indésirables.

En général, les schémas thérapeutiques décrivent des cycles qui durent trois à quatre semaines, et dont les administrations d'anticancéreux ont lieu au cours de la première semaine. Un cycle est ainsi composé d'une période dite de cure, qui définit l'intervalle de temps au cours duquel les anticancéreux sont administrés, et d'une période dite d'intercure, qui définit la période de repos, sans administration d'anticancéreux. Le schéma thérapeutique d'origine est défini méthodiquement lors de la phase 1 de son étude clinique. Mais il se diversifie souvent par la suite en multiples schémas adaptés aux réponses hétérogènes des patients à l'administration d'anticancéreux.

1.2 Les chimiothérapies dans la pratique

1.2.1 Les réponses patients-dépendantes des chimiothérapies

Les réponses aux chimiothérapies reflètent les réactions individuelles des patients aux médicaments anticancéreux. Elles regroupent les limitations de la progression du cancer d'une part, et la survenue de toxicités d'autre part.

1.2.1.1 La limitation de la progression du cancer

La limitation de la progression du cancer est l'objectif principal et souhaité du traitement par chimiothérapie. La progression par stade du cancer est décrite phénotypiquement par l'augmentation de la taille de la tumeur primitive, l'envahissement des ganglions lymphatiques et la formation de métastases. La classification TNM (Tumeur, Nœud, Métastase) [172] fournit une description détaillée de la progression du cancer à partir de ces trois phénomènes notés T, N et M.

- **T** pour Tumeur, décrit la taille et l'extension locale de la tumeur primitive. Ce paramètre peut prendre quatres valeurs de T1 à T4, indiquant une augmentation progressive de la taille ou de l'invasion locale de la tumeur.
- **N** pour Nœud, indique l'absence ou la présence d'envahissement des ganglions lymphatiques. Ce paramètre peut prendre quatre valeurs de N0 à N3, reflétant différents niveaux d'augmentation du nombre et de la localisation des ganglions atteints.
- **M** pour Métastase, indique l'absence (M0) ou la présence (M1) de métastases à distance de la tumeur primitive.

Les réponses attendues et recherchées par l'administration de chimiothérapie sont de limiter ces trois phénomènes, en rétrogradant ou en stabilisant les niveaux de la classification TNM. Mais l'une des complexités des chimiothérapies réside dans la gestion des réponses indésirables, les survenues de toxicités.

1.2.1.2 Les toxicités chimio-induites

Les toxicités associées aux chimiothérapies, également appelées chimio-induites, sont variées en terme de localisation, de forme, et de gravité. Ces toxicités sont souvent dénommées en fonction des organes affectés, telles que les cardiotoxicités, néphrotoxicités et neurotoxicités. Par exemple, les cardiotoxicités incluent des conditions comme la cardiomyopathie ou l'insuffisance cardiaque, souvent causées par des anthracyclines. Les néphrotoxicités, causées par des agents comme le cisplatine, peuvent entraîner une insuffisance rénale aiguë ou chronique. Les neurotoxicités, telles que les neuropathies périphériques, sont fréquemment associées aux taxanes et aux sels de platine [118].

Pour décrire et quantifier la sévérité des toxicités, les oncologues utilisent le *Common Terminology Criteria for Adverse Events* (CTCAE) [31]. Cette classification regroupe les toxicités selon des critères physiologiques, anatomiques, étiologiques ou des résultats cliniques. Chaque terme du CTCAE représente un événement de toxicité spécifique et est accompagné de définitions en langage naturel pour en préciser la signification. L'échelle de grades du CTCAE est divisée en cinq niveaux de sévérité :

- Grade 1 : Léger ; asymptomatique ou symptômes légers ; diagnostic à l'examen clinique uniquement ; ne nécessitant pas de traitement.
- Grade 2 : Modéré ; nécessitant un traitement minimal, local ou non-invasif ; interférant avec les activités de la vie quotidienne.
- Grade 3 : Sévère ou médicalement significatif mais sans mise en jeu immédiate du pronostic vital ; indication d'hospitalisation ; invalidant ; interférant avec les activités élémentaires de la vie quotidienne.
- Grade 4 : Mise en jeu du pronostic vital ; nécessitant une prise en charge en urgence.
- Grade 5 : Décès lié à l'événement indésirable.

L'utilisation des grades permet de quantifier et de comparer la sévérité des toxicités de manière standardisée, offrant ainsi une base solide pour l'analyse des données cliniques et la prise de décision thérapeutique. Le CTCAE est un outil essentiel dans le suivi des réponses aux chimiothérapies, permettant aux praticiens de normaliser, surveiller et adapter les traitements en fonction des réactions spécifiques des patients. En intégrant certains éléments de la terminologie *Medical Dictionary for Regulatory Activities* (MedDRA) [129], le CTCAE assure une interopérabilité accrue avec d'autres systèmes de pharmacovigilance, facilitant ainsi la communication et l'analyse des données sur les effets indésirables des traitements anticancéreux. Le CTCAE constitue également une ressource précieuse pour la représentation des connaissances sur les toxicités, comme nous le verrons dans les sections 2.3.4.3 du chapitre 2 et 4.2.2.1 du chapitre 4.

Certaines toxicités induites par la chimiothérapie peuvent être irréversibles. TRENDOWSKI et al. [204] ont montré que le cisplatine est associé à une néphrotoxicité, une ototoxicité et une neurotoxicité sévères, qui peuvent persister même après l'arrêt du traitement. Ces toxicités peuvent avoir des effets dévastateurs sur la qualité de vie des patients et sont souvent difficiles à gérer.

Enfin, il est parfois difficile de déterminer si une toxicité est due à la chimiothérapie ou au cancer lui-même. De nombreux symptômes rapportés par les patients, comme la fatigue, la perte d'appétit et la douleur, peuvent être attribués soit à la progression du cancer, soit aux effets indésirables des traitements. Cette difficulté est accentuée par le chevauchement des symptômes entre la maladie et les traitements, ce qui complique la prise de décision clinique [211].

TRENDOWSKI et al. [204] ont également montré que la survenue de certaines toxicités irreversibles étaient corrélée à l'âge et à un polymorphisme génétique. Les réponses aux chimiothérapies, sont en effet très patient-dépendantes. Nous l'illustrons dans la section suivante.

1.2.1.3 L'hétérogénéité des réponses en fonction des caractéristiques des patients

L'hétérogénéité de ces réponses est fortement influencée par de nombreux facteurs propres au patient, comme l'âge [147], les comorbidités [191], et le profil génétique [167, 56].

L'âge des patients est souvent associé à des réactions toxiques accrues et une efficacité réduite du traitement. BALDUCCI et EXTERMANN [13] ont observé que les patients de plus

de 70 ans souffrant de leucémie myéloïde, de lymphomes à grandes cellules et de carcinome cœlomique de l'ovaire présentent des réponses moins favorables aux chimiothérapies, avec une incidence accrue de cardiotoxicités et de neurotoxicités. Par contraste, CASCINU, DEL FERRO et CATALANO [30] n'ont pas identifié de différence significative d'efficacité entre les patients âgés et les plus jeunes, et soutiennent l'utilisation de schémas thérapeutiques identiques pour ces deux groupes. Chez les patients atteints du cancer du poumon non à petites cellules, DEPPERMAN [51] a observé que les patients âgés traités avec la vinorélbine avaient une survie significativement meilleure par rapport à ceux qui n'ont pas reçu ce traitement en raison de leur âge avancé. Sur les cancers du pancréas avancés traités par gemcitabine, NAKAI et al. [138] ont démontré que ce sont les comorbidités, et non l'âge en lui-même, qui affectent négativement la réponse aux chimiothérapies. Cela justifie ainsi l'administration de ce traitement chez les patients âgés sans comorbidité relevée. La corrélation entre l'âge et les comorbidités complique parfois l'identification des facteurs influençant la réponse aux traitements. Les schémas thérapeutiques sont souvent adaptés pour les patients âgés et/ou atteints de comorbidités [168, 111].

Les polymorphismes génétiques sont également à l'origine de variabilité dans les réponses à certaines chimiothérapies. Nous l'illustrons avec trois exemples connus de la littérature : les polymorphismes des gènes codant pour les enzymes DPD (dihydropyrimidine dehydrogenase), TPMT (thiopurine methyltransferase) et UGT1A1 (UDP glucuronosyltransferase famille 1 membre A1). Certains polymorphismes ou certains variants du gène DPYD peuvent causer une déficience de l'enzyme DPD qui catalyse le 5-fluorouracil (5-FU) et la capecitabine, deux anticancéreux communément administrés dans les chimiothérapies du cancer du colon. L'administration de ces molécules chez les patients déficients en enzyme DPD cause de graves neutropénies et une réponse moins efficace [208]. Certain allèles du gène TPMT réduisent l'activité enzymatique TPMT et engendrent des myélodépressions chez les patients atteints de leucémies ou de maladies auto-immunes et recevant des chimiothérapies à base d'anti-cancéreux de la classe des thiopurines [193]. Enfin les variants du gène UGT1A1 sont à l'origine de la variabilité des réponses et notamment la survenue de neutropénie chez les patients atteints de cancers du colon ou du poumon à petites cellules et recevant des chimiothérapies à base d'irinotecan [196]. Les profils génétiques des patients candidats à ces chimiothérapies sont étudiés et les schémas thérapeutiques adaptés [122].

On a mis en évidence l'importance de la prise en compte des caractéristiques individuelles des patients pour la sélection d'un schéma thérapeutique adéquat. Cependant, les réponses aux chimiothérapies ne sont pas statiques ; elles évoluent au cours du temps. Des ajustements dynamiques sont souvent nécessaires pour optimiser continuellement l'efficacité du traitement et minimiser les effets secondaires.

1.2.2 Le déroulement des chimiothérapies

Lorsqu'un cancer est diagnostiqué, différents experts médicaux se réunissent lors d'une concertation multidisciplinaire (RCP) et déterminent ensemble du traitement ou de la combinaison de traitements appropriés en fonction :

- des caractéristiques individuelles du patient et de ses préférences (*cf. section 1.2.1*),
- du stade, du type et de la localisation de la tumeur. (*cf. section 1.1.1*)
- des objectifs de traitement. (*cf. 1.1.2*)

Si le traitement choisi inclut une chimiothérapie, les experts médicaux sélectionnent un schéma thérapeutique adéquat. De nombreux guides de référence existent pour les recommandations de schémas thérapeutiques. Parmi eux, on peut citer le manuel de DeVita [37], et au niveau international, les recommandations de l'Organisation Mondiale de la Santé (OMS) [215]. En Europe, on peut citer celles de la Société Européenne d'Oncologie Médicale [62] aux États-Unis, celles du *National Comprehensive Cancer Network* [141]. Au Royaume-Uni, le *National Institute for Health and Care Excellence* [142] fournissent des recommandations spécifiques, et en France, la Haute Autorité de Santé (HAS) [85] ainsi que l'INCa [100] en fournissent également. La diversité de ces sources soulève inévitablement des questions quant à leur alignement et leur harmonisation.

Une fois le schéma thérapeutique sélectionné, le protocole de chimiothérapie, théorique et prévu, consiste en la répétition du cycle décrit par le schéma thérapeutique. La communauté oncologique utilise communément le terme de "ligne de chimiothérapie" pour qualifier l'enchaînement de cycles réellement suivis par le patient. Le nombre de cycles d'une ligne varie en fonction des réponses du patients qui sont évaluées par l'équipe médicale tout au long du traitement. Si le patient supporte mal son traitement ou que les réponses sont jugées insuffisamment efficaces, l'équipe d'experts médicaux se réunit pour une nouvelle RCP et adapte le schéma thérapeutique : le patient suit alors une nouvelle ligne de chimiothérapie.

On clarifie ici l'usage des termes "schéma thérapeutiques", "protocole" et "ligne" de chimiothérapies dans cette thèse.

- Un **schéma thérapeutique** est la description d'un cycle de chimiothérapie, avec sa durée, sa combinaison de médicaments anti-cancéreux et leurs paramètres d'administration (dose, jour, voie d'administration et mode d'infusion, introduits dans la sous-section 1.1.4).
- Un **protocole** correspond au plan thérapeutique théorique et prévu, c'est-à-dire la répétition du cycle décrit par le schéma thérapeutique.
- Une **ligne** de traitement correspond à l'enchaînement des cycles réellement suivis par le patient.

Nous avons décrit les étapes d'initiation et le déroulement d'une chimiothérapie, et souligné à nouveau le caractère évolutif des réponses aux chimiothérapies. Ces réponses sont évaluées régulièrement par l'équipe de soin, et nous allons voir comment.

1.2.3 Le suivi des réponses aux chimiothérapies

Tout au long des chimiothérapies, les soignants évaluent l'efficacité et l'innocuité des traitements au travers de consultations, de bilans sanguins et d'examens d'imagerie. Les consultations avec le patient sont essentielles pour s'assurer de son état général et pour prendre en compte son avis pour la suite du traitement. Il existe de nombreux scores et classifications pour aider les médecins à évaluer les réponses du patient au traitement. Pour évaluer la sévérité des effets indésirables, ils ont recours aux critères définis par le CTCAE qui classifient la sévérité de ces effets en cinq grades. Les échelles de Karnofsky [178] et Zubrod [225] et le score ECOG [153] se concentrent sur l'adéquation entre la qualité de vie du patient et le suivi de sa chimiothérapie. La classification TNM et les critères RECIST (*Res-*

ponse *Evaluation Criteria in Solid Tumors*) [156] se focalisent eux plus sur l'évaluation de l'efficacité du traitement, c'est-à-dire la réduction des tumeurs primitives ou métastatiques.

1.2.4 L'adaptation des chimiothérapies aux réponses des patients

Suite à ces évaluations, l'oncologue peut décider modifier le schéma thérapeutique en réduisant les doses ou le nombre de jours entre certaines administrations d'anti-cancéreux. Si le traitement est jugé trop inefficace ou nocif, les experts médicaux se réunissent pour une nouvelle RCP pour changer le schéma thérapeutique. Le patient suit alors une nouvelle ligne de traitement.

1.2.5 L'adhésion aux chimiothérapies

1.2.5.1 Définition

Selon l'OMS, l'adhésion thérapeutique est définie comme « le degré d'adéquation du comportement d'une personne – prise de médicaments, respect d'un régime alimentaire et/ou modification du mode de vie – avec les recommandations convenues avec un professionnel de santé » [155]. Cette définition met en avant une collaboration active entre le patient et les professionnels de santé pour gérer le traitement, ce qui distingue l'adhésion de la notion de compliance et de l'observance, ces dernières impliquant un suivi passif des prescriptions sans intégration active du patient.

1.2.5.2 Justification de l'utilisation du terme “adhésion” pour les chimiothérapies non orales

Dans le contexte des chimiothérapies administrées par voies non orales (telles que les injections intraveineuses ou intramusculaires), le terme “adhésion” est particulièrement pertinent. Ces traitements nécessitent la présence de professionnels de santé pour leur administration, ce qui assure une surveillance et un suivi médical rigoureux du patient. Cette dynamique ne se limite pas seulement à l'administration correcte des traitements mais inclut également une adaptation continue du plan thérapeutique en fonction des réponses du patient, des effets secondaires observés et des préférences personnelles du patient. L'utilisation du terme “adhésion” reflète donc l'engagement actif du patient dans le processus de traitement, en collaboration avec l'équipe soignante, pour optimiser les résultats thérapeutiques tout en tenant compte de son mode de vie et de ses valeurs personnelles.

1.2.5.3 Facteurs qui contribuent à l'adhésion incomplète au sein des lignes de chimiothérapies

Plusieurs facteurs peuvent conduire à une adhésion incomplète :

- **Les stratégies médicales :** Les médecins adaptent souvent le traitement en fonction de la réponse au traitement et de l'état général du patient, en ajustant les doses ou en modifiant le calendrier des administrations, c'est-à-dire les retards de cycles ou décalage de jours d'administration (*cf. sous-sections 1.2.3 et 1.2.4*)
- **La commodité pour le patient :** Des adaptations peuvent être accordées pour s'adapter au mode de vie du patient. Il est par exemple fréquent d'effectuer une pause estivale pour limiter les effets secondaires du patient pendant ses vacances.
- **Des contraintes opérationnelles des établissements de santé :** Les planifications et la disponibilité du personnel, ainsi que les contraintes économiques de l'établissement

ment, peuvent influencer l'adhésion au traitement planifié.

Quelle que soit leurs raisons, les modifications apportées aux protocoles de chimiothérapie, sont le résultat d'une évaluation médicale considérant l'impact acceptable sur le pronostic du patient. Toutefois, l'impact réel de ces ajustements sur les résultats cliniques reste largement méconnu et fait l'objet de divergences d'opinions parmi les oncologues. Pour de nombreux schémas thérapeutiques, il y a des débats actifs au sein de la communauté médicale, concernant l'équilibre entre le respect strict du protocole standard et la personnalisation du traitement. Les incertitudes quant aux meilleures pratiques nous conduisent à explorer les adaptations praticien-dépendantes ou organisation-dépendante, illustrées par des exemples concrets dans la littérature médicale dans la section suivante.

1.2.6 Les adaptations praticien-dépendantes aux protocoles

Nous avons mentionné les stratégies médicales parmi les facteurs de l'adhésion incomplète au traitement (*cf.* section 1.2.5.3). Les adaptations aux protocoles de chimiothérapie peuvent varier selon les praticiens, en raison de nombreux facteurs tels que l'expérience clinique et les interprétations des études scientifiques.

Une illustration frappante de cette variabilité est observée dans l'administration de 5-fluorouracil (5-FU), un agent chimiothérapeutique largement utilisé dans le traitement du cancer colorectal. Malgré des décennies d'utilisation, il n'existe aujourd'hui pas de consensus clair parmi les oncologues quant au meilleur mode d'administration du 5-FU : en bolus ou en infusion continue. Le bolus correspond à un mode d'administration avec une dose élevée et une durée d'infusion rapide. Dès 1975, SEIFERT et al. [184] comparent les infusions lente et en bolus du 5-FU avec un essai randomisé et montrent que l'infusion continue du 5-FU améliore la tolérance sans compromettre l'efficacité du traitement. Ce résultat est confirmé par une étude de PIEDBOIS et al. [160] menée en 1998. Malgré ces études, les schémas thérapeutiques de Gramont [77] qui combinent administrations en bolus et continue du 5-FU, sont couramment utilisés pour traiter les cancer colorectaux. Une revue de la littérature parue en 2022 [6] montre que des résultats de nombreuses études sont en faveur de l'omission de l'administration du 5-FU en bolus pour réduire la survenue toxicités sans réduire l'efficacité des traitements. Toutefois, l'étude conclut qu'un essai randomisé est nécessaire pour clore le débat sur l'utilité de l'administration du 5-FU en bolus.

Un autre exemple de désaccord notable concerne l'administration du carboplatine dans le traitement du cancer du poumon non à petites cellules (NSCLC) chez les patients âgés. La dose optimale de carboplatine pour cette population reste un sujet de débat parmi les oncologues. Certains praticiens préconisent une réduction de la dose pour les patients de plus de 70 ans en raison d'une tolérance généralement moindre aux effets secondaires de la chimiothérapie chez les personnes âgées. Par exemple, la recherche menée par QUOIX et al. [164] a montré que les doses réduites étaient mieux tolérées et associées à une meilleure qualité de vie. Cependant, d'autres experts soutiennent que la réduction systématique des doses peut compromettre l'efficacité du traitement. Une étude de GRIDELLI et al. [80] souligne que, pour certains patients en bonne condition physique, des doses standard peuvent offrir de meilleurs résultats en termes de survie sans progression de la maladie, même chez les personnes âgées.

Enfin, un dernier exemple de débat concerne l'utilisation du bévacizumab, un anticorps monoclonal, dans le traitement du cancer colorectal métastatique. Le principal point de désaccord réside dans la durée optimale du traitement par bévacizumab. Certaines études, comme celle de SALTZ et al. [176], suggèrent que le bévacizumab devrait être administré de manière continue tant que la maladie n'a pas progressé, et montrent une meilleure survie sans progression. En revanche, d'autres recherches, telles que l'étude de KABBINAVAR et al. [107], montrent que l'ajout de bévacizumab améliore les résultats globaux mais ne clarifient pas la durée idéale du traitement. Par ailleurs, certains cliniciens s'inquiètent des coûts élevés et des risques accusés d'effets secondaires graves, tels que l'hypertension et les complications thrombotiques, associés à une utilisation prolongée du bévacizumab. Ces préoccupations ont conduit à des pratiques variées, certains oncologues optant pour une durée de traitement fixe et limitée, tandis que d'autres adaptent la durée en fonction de la réponse individuelle du patient et de sa tolérance au traitement.

Ces exemples illustrent comment l'absence de consensus peut influencer les décisions thérapeutiques, et conduire à des adaptations des schémas thérapeutiques selon les hypothèses, l'expérience et l'environnement de travail des praticiens. Ils soulignent le besoin de lignes directrices plus claires.

1.3 Évaluation des réponses aux chimiothérapies dans les études cliniques

1.3.1 Évaluation des schémas thérapeutiques

À l'initiation d'un traitement de chimiothérapie, il est parfois proposé au patient de participer à un essai thérapeutique. Ces essais ont pour but d'évaluer l'efficacité d'un nouveau schéma thérapeutique par rapport à un schéma thérapeutique standard, ou de deux options de schémas thérapeutiques standards [92, 23]. Les deux bras de l'essai correspondent aux deux groupes de patients suivant les différents schémas. Pour évaluer l'efficacité des schémas, les chercheurs comparent la survie globale SG (ou l'*Overall Survival* - OS en anglais) ou la survie sans progression SSP (*Progression Free Survival* - PFS) des deux branches. L'analyse de survie est une branche de la statistique qui modélise la durée jusqu'à la survenue d'un événement (cf. section 3.2.1 du chapitre 3). En oncologie, la survie sans récidive mesure le temps jusqu'à la rechute du cancer, et la survie globale mesure le temps jusqu'au décès du patient. Dans ces études, un schéma thérapeutique est jugé plus efficace qu'un autre s'il existe une différence significative entre les survies des patients.

1.3.2 Évaluation de l'adhésion aux schémas thérapeutiques

De nombreuses études rétrospectives cherchent à montrer l'impact de réduction de dose et des retards dans les schémas thérapeutiques. Dans ces études, on cherche à savoir si la réduction de dose d'un ou plusieurs anticancéreux a un impact sur la survie des patients. Il existe des objectifs multiples et variés à ces études, qui dépendent des schémas thérapeutiques et localisations de cancers. La plupart des études cherchent à montrer qu'il est important d'adhérer au maximum aux schémas thérapeutiques, de ne pas réduire les doses des administrations ni de les décaler dans le temps [121]. Certaines montrent au contraire que la réduction d'anticancéreux n'a pas d'impact significatif sur la survie des patients.

âgés et améliorent leur qualité de vie [82]. Quelque soit l'objectif, la méthode traditionnelle pour mesurer l'adhésion au protocole est le calcul de la dose-intensité relative définie par HRYNIUK [91] en 1988.

Définition - Dose-intensité (DI)

Soit d un médicament anti-cancéreux. Sa dose intensité DI est définie comme

$$\text{DI}(d) = \frac{\text{Dose totale reçue du médicament } d}{\text{durée en semaines depuis le début du traitement}}. \quad (1.1)$$

Définition - Dose-intensité relative (RDI)

La dose-intensité relative, notée RDI, est le ratio de la dose d'anti-cancéreux reçue par le patient sur la dose qu'il aurait reçu en suivant le protocole [120]

$$\text{RDI}(d) = \frac{\text{DI}(d) \text{ reçue}}{\text{DI}(d) \text{ du protocole}}. \quad (1.2)$$

Définition - Dose-intensité relative pour tous les médicaments (ADRD)

La dose-intensité relative pour tous les médicaments anti-cancéreux d'un protocole est notée ADRDI, pour *All Drug Relative Dose Intensity*. Il s'agit de la moyenne arithmétique de la dose-intensité relative de chaque médicaments anti-cancéreux. En notant \mathcal{M} l'ensemble des médicament anticancéreux d'un protocole, elle s'écrit

$$\text{ADRD} = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \text{RDI}(d_i) \quad (1.3)$$

La plupart du temps, les chercheurs fixent un seuil de dose intensité relative, souvent aux alentours de 80%, et comparent la survie des groupes de patients ayant une dose intensité relative au-dessus et en-dessous de ce seuil [221, 197, 27, 146, 50, 121].

Bien que HRYNIUK [91], DENDULURI et al. [50] et LONGO et al. [120] soutiennent que la RDI est utile pour mesurer les retards d'administration dans les chimiothérapies, les études visant à mesurer l'impact de ces retards sur la survie ne s'en servent pas forcément [222, 177]. L'impact sur la survie des ajustements temporels des protocoles de chimiothérapie restent peu connus.

1.3.3 La nécessité d'un consensus sur les schémas thérapeutiques et le potentiel des DPI

Nous avons mis en évidence la multiplicité des hypothèses sur les schémas thérapeutiques et leurs adaptations. La grande variété de ces hypothèses illustre un besoin de consensus pour améliorer la gestion des soins des patients sous chimiothérapie. De nombreux chercheurs soulignent le besoin de preuves de haut niveau à propos des recommandations

sur les protocoles de chimiothérapies [165, 78].

La généralisation des Dossiers Patients Informatisés (DPI) permet de capturer des données de vie réelle et de proposer des études observationnelles en oncologie. Le terme observationnelle est utilisé pour souligner le fait que les données ne sont pas capturés dans l'objet de l'étude, mais dans un cadre plus général, ici celui du soin. L'un des objectifs de cette thèse est de tester le potentiel des DPI à générer des preuves de haut niveau, notamment sur les recommandations pour les protocoles de chimiothérapies.

Dans la suite de ce chapitre, nous allons présenter les entrepôts de données de santé et la capture d'éléments d'information concernant les réponses aux chimiothérapies dans ces entrepôts.

1.4 Les entrepôts de données de santé

Cette section s'appuie entre autres sur les thèses de Nicolas Garcelon [70] et de Sébastien Cossin [41].

1.4.1 Les données de santé et les Dossiers Patients Informatisés

Les données de santé comprennent toutes les informations collectées sur les patients dans le cadre de leur prise en charge médicale. L'informatisation des système de santé permet l'enregistrement de données sur l'état de santé du patient. La volonté de mise en commun de ces informations pour faciliter les soins a abouti à la création du Dossier Patient Informatisé (DPI). Il s'agit d'un système centralisé de collecte et de gestion des informations médicales d'un patient. En France, l'informatisation s'est accélérée notamment pour évaluer les coûts et les remboursements des soins avec le Programme de Médicalisation des Systèmes d'Information (PMSI). Ces données médico-administratives font partie des nombreuses sources d'information du DPI.

1.4.2 La provenance et l'hétérogénéité des données des DPI

Les données médicales qui constituent le DPI proviennent de diverses sources au sein des système d'information hospitalier (SIH). Ces données sont de nature et de formats variés, et produites par différents acteurs du parcours de soin des patients. Il peut s'agir par exemple de résultats d'analyses sanguines sous format numériques générés par des laboratoire d'analyses biomédicales, d'images de scanners générés par le service de radiologie, de codes diagnostics de la Classification Internationale des Maladies (CIM10) générés par le PMSI, d'items de questionnaire saisis par des infirmiers, ou encore de comptes-rendus textuels de consultation de médecine générale. Chacune des information du DPI constitue une pièce d'un puzzle représentant le profil et l'état de santé d'un patient. À l'image du puzzle, les pièces peuvent être similaires et augmenter la confiance dans une information (*p.ex.* la mention d'une créatinine élevée à la fois dans un résultat de laboratoire et dans un compte rendu) ou être complémentaires et reconstituer le diagnostic ou la prise de décision médicale (*p.ex.* la réduction d'un anticancéreux dans le logiciel de prescription à cause d'une créatinine élevée). Les DPI produits par les hôpitaux sont alors un support intéressant pour la recherche médicale. Pour permettre leur réutilisation à des fins de recherche, il faut les stocker au sein de bases de données capables d'intégrer de gros volumes de données

variées, les entrepôts de données hospitaliers.

1.4.3 L'alimentation des entrepôts de données hospitaliers

Un entrepôt de données hospitalier est conçu pour collecter et intégrer l'ensemble des données produites lors de la prise en charge du patient à l'hôpital. Le SIH collecte et gère les données propres à chaque service, puis un processus d'extraction, transformation et chargement (Extraction Transformation Loading - ETL), souvent adapté au service, est utilisé pour intégrer les données produites à l'entrepôt de données. Ce processus garantit que les données sont correctement formatées, transformées et adaptées au schéma de l'entrepôt de données et qu'il est ainsi possible d'exploiter l'entrepôt pour les interroger. Parmi les schémas d'entrepôt de données hospitalier, nous détaillons celui d'i2b2, utilisé par les entrepôts de données de l'HEGP et du CHU de Bordeaux utilisés au cours de cette thèse.

1.4.4 Le schéma en étoile d'i2b2

Le modèle de données d'i2b2 (Informatics for Integrating Biology and the Bedside) [94] est composé de six tables et structuré autour d'une table centrale appelée table des faits (OBSERVATION_FACT), qui contient toutes les observations atomiques des patients. L'observation correspond à l'information d'intérêt. Il peut s'agir d'un code diagnostic CIM10, d'une valeur numérique de résultat de laboratoire ou du contenu textuel d'un compte rendu, etc... Cette observation est liée à un identifiant de patient, un numéro de visite pendant laquelle elle a eu lieu, un concept décrivant sa nature, le service qui l'a produit, et une date correspondant à son édition. Notez que la date correspond à une méta-donnée, l'édition de l'observation, et que cette date ne correspond donc pas forcément à la date de l'observation elle-même.

1.4.5 Les entrepôts de données de l'HEGP et du CHU de Bordeaux

L'Hôpital Européen Georges-Pompidou (HEGP), inauguré en 2000, est un établissement de santé universitaire situé dans le sud-ouest de Paris. Cet hôpital de 700 lits, et employant environ 3 000 personnes, a été le premier de l'Assistance Publique - Hôpitaux de Paris (AP-HP) à intégrer un Système d'Information Hospitalier (SIH) performant dès sa création. En 2008, l'HEGP a développé un entrepôt de données cliniques en utilisant la schéma i2b2.

L'entrepôt de l'HEGP intègre les données provenant de diverses sources du SIH, incluant notamment les résultats de laboratoires, les prescriptions médicales, les observations cliniques, les codes CIM-10, les actes médicaux, et les compte-rendus médicaux. En quelques années, cet entrepôt a consolidé les données de plus de 800 000 patients, facilitant ainsi la réalisation de nombreux projets de recherche clinique et épidémiologique. La sécurité et la confidentialité des données sont garanties par le pseudonymisation des informations identifiantes des patients [224]. L'intégration de l'entrepôt dans le SIH a permis de mener 74 projets de recherche entre 2011 et 2015, couvrant divers domaines de la santé [102]. L'entrepôt de l'HEGP constitue notre principale source de données pour le développement des projets des chapitre 4 et 5.

Le Centre Hospitalo-Universitaire (CHU) de Bordeaux s'est doté, depuis 2017 d'un entrepôt de données concentrant les informations de son SIH depuis 2010, stockant les DPI de plus de 2 millions de patients. Les similitudes techniques entre les entrepôts du CHU de

Bordeaux et de l'HEGP nous ont permis de déployer la représentation des chimiothérapies avec l'outil ChemoOnto (*cf.* chapitre 4) pour reproduire les résultats observés à l'HEGP dans un hôpital différent (*cf* chapitre 5). Nous avons notamment échangé avec l'équipe **UIAM** de Bordeaux pour qu'il déploient l'ETL présent à l'HEGP qui permet d'intégrer à l'entrepôt les enregistrements sur les chimiothérapies, que nous détaillons dans la section suivante.

1.4.6 Les enregistrements sur les chimiothérapies

1.4.6.1 Le logiciel Chimio®

À l'HEGP, comme dans 97% des établissement habilités à faire de la chimiothérapie en France, le logiciel **Chimio®** [61] gère le circuit complet des chimiothérapies, de la prescription à l'administration. Ce logiciel assure une traçabilité totale à chaque étape du processus de soin en oncologie. Le logiciel Chimio dispose de sa propre bases de données. Tous les champs de saisie du logiciel y sont intégrés dans des tables spécifiques.

Les données sur les schémas thérapeutiques

Les deux entrepôts de données d'étude suivent plusieurs schémas. Le schéma principal est celui d'i2b2 qui concentre les données cliniques. Mais il existe d'autres schémas qui intègrent des données non cliniques. Parmi eux, le schéma CHIMIO duplique quatre tables (PROTOCOLE, LIGNEPRO, PRODUIT, DCI) de la base de données du logiciel Chimio qui fournit des données sur les schémas thérapeutiques. Les tables DCI et PRODUIT concernent les médicaments (anti-cancéreux et anti-effet indésirables), la table PROTOCOLE contient des informations sur les schémas thérapeutiques (une entrée par schéma thérapeutique), la table LIGNEPRO détaille les modalités d'administration de chaque schéma thérapeutique (une entrée par administration).

Les données cliniques sur les lignes de traitement suivies

Un ETL des données cliniques entrées dans le logiciel Chimio enregistre chaque semaine les champs saisis sur les administrations médicamenteuses, notamment la dose totale administrée et le nom du médicament, mais aussi des informations sur le patient comme son poids et sa taille et le taux de créatinine, parfois nécessaires pour calculer la dose à administrer. Une chaîne de caractères avec des séparateur contenant l'intégralité des champs saisis dans Chimio pour une administration constitue une entrée dans la table OBSERVATION_FACT.

Les enregistrements du logiciel Chimio constituent une excellente source de connaissances sur les traitements de chimiothérapies suivies et théoriques. De plus, ces données sont structurées, c'est-à-dire qu'elles ne nécessitent pas ou peu de transformations pour être traitées. Leur format standard facilite leur accès à l'aide de requêtes SQL et ainsi leur analyse. Les enregistrements de réponses aux chimiothérapies sont quant à eux dispersés dans diverses sources, structurées, semi-structurées et non structurées, comme nous allons le voir dans la section suivante.

1.4.7 Les enregistrements de réponses aux chimiothérapies

Au sein de l'entrepôt de l'HEGP, nous avons identifié trois sources contenant des informations sur la réponse aux chimiothérapies. Nous détaillons celles-ci dans cette section.

1.4.7.1 Les éléments de réponses aux chimiothérapies dans les comptes-rendus hospitaliers

La richesse en information dans le texte libre Malgré les efforts visant à encourager les cliniciens à coder les informations médicales dans le DPI pour les rendre exploitables pour la recherche, une grande partie des informations demeure souvent sous forme de texte libre [66, 145]. Plusieurs raisons expliquent cette situation :

- la rédaction narrative de l'histoire du patient permet aux cliniciens d'exprimer leurs doutes et leur raisonnement clinique de manière détaillée, ce qui n'est pas possible avec les données codées
- il est souvent plus rapide et moins contraignant de saisir directement les informations dans un champ de texte libre plutôt que de chercher la case à cocher appropriée, le temps passé avec le patient étant précieux [203]
- le texte libre permet d'enregistrer des détails qui peuvent sembler insignifiants aujourd'hui, mais qui pourraient s'avérer importants pour comprendre l'évolution d'une maladie à l'avenir. Réduire la description du patient à des éléments codés peut entraîner une perte significative d'informations cruciales pour la découverte de nouvelles connaissances.

En oncologie, les réponses aux chimiothérapies sont fréquemment décrites dans le texte libre. Cependant, les systèmes d'évaluations d'efficacité et de sécurité des traitements que nous avons mentionnés dans la section 1.2.3, tels que la classification TNM ou les critères du CTCAE, sont également inclus dans ces textes libres. Ces évaluations sont précieuses car elles sont directement employées pour décrire les réponses aux traitements de façon spécifique et sont pour cela plus facilement identifiables que des descriptions narratives.

Dans les parties semi-structurées

Certains services utilisent des modèles de compte rendu avec des sections. Ceci permet d'associer un contexte aux différentes parties du texte et simplifie l'interprétation des informations extraites automatiquement. Ces modèles de comptes-rendus contiennent parfois aussi des tableaux, permettant là encore une facilité de l'extraction et l'interprétation des informations. Notez que les modèles de ces comptes-rendus sont propres au service de l'hôpital. Les processus d'extraction d'information sont donc difficilement généralisables à d'autres hôpitaux. Au sein de l'entrepôt de l'HEGP, nous avons identifié un modèle de compte rendu contenant un tableau avec des informations sur les toxicités et leurs grades.

1.4.7.2 La collecte des toxicités dans les questionnaires

Dans le cadre de certaines prises en charge, les patients peuvent être interrogés la veille d'un cycle de chimiothérapie, sur les toxicités survenues au cours de leur cycle actuel. À l'HEGP, nous avons identifié deux types de questionnaires collectant ces toxicités, le questionnaire "Proche" et le questionnaire "Anticip". Les items de ces questionnaires sont constituées du nom des toxicités, et la réponse au questionnaire correspond au grade associé. Ces questionnaires sont des données structurées, de haute qualité et facilement analysables.

1.4.8 Les données de décès de l’Insee

Depuis 2021, l’Institut National de la Statistique et des Études Économiques (l’Insee) a rendu public la base nationale des personnes décédées depuis 1970. Les entrepôts de l’HEGP et du CHU de Bordeaux ont tous les deux intégré ces dates de décès à leur entrepôts, via des algorithmes différents de mise en correspondance entre patients et personnes du fichier de décès de l’Insee [41]. Ces dates de décès sont nécessaires pour mesurer la survie globale des patients, et évaluer les réponses aux chimiothérapies en s’inspirant de ce qui est fait dans les études cliniques (*cf.* section 1.3). Pour détecter les réponses aux chimiothérapies, nous nous inspirons en effet de ce qui est fait dans les études cliniques, avec des études de survie (*cf.* section 3 et chapitre 5). Il faut toutefois clarifier les différences entre les études cliniques classiques et celles faites à partir de la réutilisation de DPI.

1.5 Les défis à relever pour réutiliser les DPI à des fins de recherche

1.5.1 Données dédiées à la recherche vs. données de vie réelles réutilisées pour la recherche

Les données de vie réelle, qui constituent les DPI, et les données cliniques récoltées spécifiquement dans le cadre d’études dédiées ne sont pas collectées dans le même but. Les données des DPI sont renseignées dans le cadre de l’activité de soin, tandis que les données des essais cliniques sont collectées spécifiquement pour répondre à des hypothèses de recherche. Dans le cadre des essais clinique, le recueil quasi-systématique des données assure une bonne complétude de la base de données résultante. Les essais cliniques suivent des protocoles rigoureux pour garantir la qualité et la précision des données, alors que les DPI, même s’ils contiennent une grande richesse d’informations, peuvent présenter des problèmes de qualité ou être difficilement exploitables. Ainsi, la réutilisation des DPI pour des recherches similaires soulève plusieurs défis qu’il est essentiel de relever pour assurer la validité des résultats obtenus [46, 201].

1.5.2 La qualité des DPI et leur validité

Les DPI peuvent contenir des informations manquantes, des erreurs de saisie et une diversité de formats de données. Ces imperfections sont dues à la nature même des DPI, qui sont renseignés en routine clinique pour des objectifs de soins immédiats et non pour des recherches ultérieures. Les erreurs de saisie sont plus fréquentes dans les DPI que dans les essais cliniques, où des procédures strictes de contrôle de qualité sont mises en place pour garantir la fiabilité des données. Par conséquent, les analyses basées sur les DPI doivent inclure des étapes rigoureuses de nettoyage et de validation des données pour minimiser les biais et les erreurs [46, 201].

1.5.3 La nécessité d’extraire, de représenter et d’intégrer les informations des DPI

Les DPI constituent une mine d’informations précieuses, mais celles-ci ne sont pas toujours facilement accessibles. Il est nécessaire d’utiliser des techniques avancées de fouille

de données pour extraire les informations pertinentes tel que des outils de traitement automatique des langues (TAL) pour analyser le texte libre, et des techniques d'intégration de données pour combiner des sources hétérogènes [46] (cf. chapitre 3).

Ces techniques doivent être évaluées pour mesurer la qualité de l'information extraite. Bien que ces méthodes permettent d'accéder à une grande quantité d'informations, la fiabilité de ces données ne sera jamais aussi élevée que celle des données d'essais cliniques dont le recueil est rigoureusement contrôlé. Les analyses basées sur les DPI doivent donc être interprétées avec prudence, en tenant compte des limites inhérentes à ces sources de données.

1.5.4 Différence de conclusion et implications des résultats obtenus avec des DPI

Alors que les preuves obtenues à partir des données de vie réelle (*real-world evidence*) peuvent fournir des informations précieuses et compléter les résultats des essais cliniques, elles ne peuvent pas remplacer les essais cliniques randomisés pour établir des liens de causalité. Les résultats issus des DPI peuvent servir comme une méthode exploratoire pour générer des hypothèses, mais la validation de ces hypothèses nécessite souvent des essais cliniques supplémentaires [201].

1.6 Bilan : Les défis de la réutilisation des DPI pour étudier les réponses aux chimiothérapies

Ce premier chapitre a permis de poser les bases indispensables à la compréhension des enjeux liés aux chimiothérapies dans le traitement du cancer, tout en mettant en lumière les défis complexes que pose la réutilisation des données de santé issues des DPI.

Nous avons identifié les facteurs suivants de complexité des chimiothérapies :

- Les chimiothérapies suivent une organisation rigoureuse incluant des cycles, des lignes de traitement et une variété recommandations sur les schémas thérapeutiques. (cf. sections 1.1.4, 1.2.2)
- Les réponses aux traitements de chimiothérapie sont patient-dépendantes. Elles varient considérablement en fonction de nombreux facteurs tels que l'âge, les comorbidités et le profil génétique des patients. (cf. section 1.2.1)
- La gestion de l'efficacité et de la tolérance au traitement est une coopération entre patients et médecins, et nécessite une surveillance continue et des adaptations en conséquence. (cf. sections 1.2.3, 1.2.4, 1.2.5, 1.2.6)

Nous avons mis en évidence l'hétérogénéité et la qualité des données de santé et le besoin de développer des outils de fouille et de représentation pour les réutiliser :

- Les DPI constituent une source riche d'informations cliniques, mais leur hétérogénéité et la variabilité de leur qualité rendent difficile leur utilisation à des fins de recherche.
- Pour exploiter pleinement le potentiel des DPI, il faut développer des techniques sophistiquées d'extraction, de représentation et d'intégration des informations.

- Les méthodes traditionnelles de traitement des données ne suffisent pas pour gérer la complexité et l'ampleur des données cliniques disponibles.

Dans le chapitre suivant nous explorons comment les ontologies peuvent être utilisées pour formaliser les concepts relatifs aux chimiothérapies et à leurs réponses, pour permettre une intégration efficace des données issues des DPI. Cette exploration ouvrira la voie à une exploitation optimale des données cliniques, pour faciliter la détection et la prédiction des réponses aux traitements

Deuxième partie

État de l'art

CHAPITRE

2

REPRÉSENTATION DE CONNAISSANCES

2.1	Donnée, information, connaissance	38
2.2	Représentation de connaissances	39
2.2.1	Les représentations des connaissances dans le domaine bio-médical	39
2.2.2	Bases des logiques de descriptions	43
2.2.3	Le raisonnement automatique	44
2.2.4	Ontologies, bases de connaissances, système à base de connaissances, graphes de connaissances	44
2.3	Représenter les connaissances avec les outils du Web sémantique	46
2.3.1	Les outils du Web sémantique pour implémenter la représentation d'un domaine dans la pratique	46
2.3.2	Resource Description Framework (RDF)	47
2.3.3	Identifier les ressources sur le Web avec les URI	48
2.3.4	Des modèles de référence à réutiliser	49
2.3.5	Gérer et requêter un graphe de connaissances avec un triple store et SPARQL	51
2.4	Raisonnement et fouille de graphe	52
2.4.1	Le raisonnement dans un système à base de connaissances	52
2.4.2	Vers la fouille des graphes de connaissances	54
2.5	État de l'art des modèles de données spécifiques aux chimiothérapies	55
2.5.1	Note sur les représentations temporelles de la TEO et de la TO	57
2.6	Bilan : pourquoi les ontologies ?	58



De nombreuses études ont déjà mis en évidence le potentiel des preuves empiriques disponibles dans les entrepôts de données hospitaliers [200, 186]. Cependant, ces données demeurent sous-exploitées pour plusieurs raisons :

- La structure des entrepôts de données est parfois inadaptée pour représenter la complexité d'un domaine de santé spécifique.
- Les connaissances sont dispersées dans des sources hétérogènes au sein des entrepôts de données, rendant leur exploitation complexe.

- Les entrepôts de données ne sont pas directement connectés à d'autres modèles de connaissances.

Dans ce chapitre, nous illustrons avec différents exemples tirés de la littérature, comment l'utilisation d'ontologies permet de résoudre ces trois problèmes.

Nous commençons par définir les distinctions entre donnée, information et connaissances. (section 2.1)

Nous montrons comment ces connaissances peuvent être représentées, d'abord en exposant les représentations principalement utilisées dans le domaine bio-médicale, puis en nous concentrant sur les ontologies. Nous présentons les bases de logiques de description et le raisonnement automatique. Nous terminons la section par les définitions de bases de connaissances, graphes de connaissances et systèmes à base de connaissances. (section 2.2) Puis, nous présentons les technologies standards du Web Sémantique pour implémenter les ontologies. Nous montrons comment gérer et requêter une base de connaissances. (section 2.3)

Dans une quatrième section, nous montrons comment le raisonnement et les connexions avec différents modèles de données permet l'exploration de connaissances (section 2.4). Nous expliquons comment le raisonnement est utilisé dans les systèmes à base de connaissances (sous-section 2.4.1). Enfin nous listons quelques méthodes de fouille applicables dans les graphes de connaissances (sous-section 2.4.2).

Dans la cinquième section, nous proposons un état de l'art de la représentation de chimiothérapies et de leurs réponses. (section 2.5)

Nous terminons par un bilan pour résumer pourquoi nous pensons que les ontologies sont pertinentes pour représenter les chimiothérapies et leurs réponses. (section 2.6)

Ce chapitre s'appuie sur les thèses de Pierre Monnin [133], de Maxime Delmas [48] et du livre "Python et les ontologies" de Jean-Baptiste Lamy [103].

2.1 Donnée, information, connaissance

Dans ce chapitre, on s'intéresse à la représentation des connaissances. En introduction, nous avons mentionné que nous extrayions des connaissances à partir des données. Nous posons ici les distinctions entre les notions de "données", d'"information" et de "connaissance" en nous appuyant sur SCHREIBER [181].

Définition - Données

Les données^a sont des symboles non interprétés.

a. SCHREIBER [180]

Par exemple, un signal numérique reçu par un thermomètre est une donnée brute sans interprétation particulière.

Définition - Information

L'information^a consiste en des données dotées d'une signification permettant leur in-

terprétation par un être humain.

a. SCHREIBER [180]

Par exemple, une température de 38°C mesurée par un thermomètre est interprétée comme une indication de fièvre chez une personne.

Définition - Connaissance

La connaissance^a est définie comme des informations et des données assimilées pouvant être utilisées pour accomplir des tâches et créer de nouvelles informations. Ainsi, la connaissance est exploitable puisqu'elle peut soutenir l'accomplissement d'une tâche ou la prise de décision. Elle possède également une capacité génératrice car elle peut être utilisée pour créer de nouvelles informations. SCHREIBER [180] souligne que cette génération de nouvelles informations est l'une des fonctions majeures de la connaissance.

a. SCHREIBER [180]

Par exemple, un médecin utilise des connaissances sur les symptômes de la fièvre pour diagnostiquer une maladie et prescrire un traitement approprié.

Une connaissance peut s'exprimer par une proposition en langage naturel et énonce des faits sur la réalité. Ainsi, les compte rendus cliniques regorgent de connaissances. La capacité génératrice de la connaissance décrite par SCHREIBER [180] est liée au raisonnement humain qui déduit de nouvelles connaissances. Ces déductions sont aussi appelées des inférences [64].

2.2 Représentation de connaissances

2.2.1 Les représentations des connaissances dans le domaine bio-médical

Il existe un spectre des représentations de connaissances, allant des glossaires aux ontologies [64]. Dans le domaine bio-médical les principales représentations de connaissances utilisées sont les terminologies et les ontologies. Avant de les détailler, nous définissons quelques notions permettant de décrire un domaine de connaissances [103, 81, 209].

Définition - Concept

Un concept est une idée ou une notion abstraite qui représente une catégorie ou un type d'éléments partageant des caractéristiques communes.

Par exemple, le concept d'"œsophagite" décrit l'ensemble des inflammations de l'œsophage.

Définition - Terme

Un terme est un mot ou une expression utilisée pour désigner un concept. Plusieurs

termes différents peuvent décrire ou être englobés par le même concept. Les termes sont les éléments de base des terminologies et sont utilisés pour nommer et décrire les concepts.

Par exemple, "rétrécissement de l'œsophage", "sténose œsophagienne" et "inflammation de l'œsophage" sont tous des termes différents qui peuvent décrire le même concept d'œsophagite.

Définition - Instance

Une instance est une occurrence spécifique d'un concept dans un modèle de données. Elle représente un exemple particulier, concret et identifiable de manière unique du concept abstrait.

Par exemple, Bob, identifié par son numéro patient et atteint d'une œsophagite peptique, instancie le concept "Patient".

Définition - Relation

Une relation est un lien, une association entre deux entités (qui peuvent être des concepts ou des instances). Les relations peuvent être hiérarchiques ou transversales.

Par exemple, la relation "a un diagnostic de" pourrait relier l'instance "Bob" à l'instance "œsophagite peptique". Le concept "Patient" peut être relié par une relation de subsumption au concept "Personne" pour signifier que les patients sont des personnes.

Définition - Entité

Les entités correspondent à l'ensemble des éléments, abstraits (concepts, relations possibles entre les instances de deux concepts) ou concrets (instances de concepts et relations), d'un domaine que l'on cherche à représenter avec un modèle de données. Elles englobent les concepts ainsi que les instances spécifiques de ces concepts.

Par exemple, une entité peut être la mention de l'occurrence d'une "œsophagite peptique" dans un compte-rendu, ou un certain patient, Bob, identifié par son numéro patient et atteint de cette inflammation, ou encore le concept d'œsophagite.

Pour représenter des domaines spécifiques, les communautés bio-médicales utilisent principalement deux représentations de connaissances :

Définition - Terminologies

Les terminologies sont des ensembles de termes et de définitions spécifiques à un domaine, souvent organisés de manière hiérarchique. Elles incluent :

- Les **thésaurus** : répertoires de termes interconnectés par des relations hiérarchiques, synonymiques et associatives.

- Les **classifications** : systèmes d'organisation de classes selon des critères définis.

Par exemple, le *National Cancer Institute Thésaurus* (**NCIt**) est un répertoire de termes en oncologie. La *Common Terminology Criteria for Adverse Events* (**CTCAE**) [31], que nous avons mentionnée lors du chapitre 1, est une classification qui utilise des critères pour classer les toxicités en cinq grades selon leur sévérité.

Définition - Ontologies

Les ontologies sont des descriptions formelles et explicites de l'ensemble des concepts d'un domaine et des relations existant entre les entités de ces concepts. Ces relations entre concepts peuvent être hiérarchiques et transversales.

Les ontologies sont des descriptions "formelles" car interprétables par une machine, et "explicite" car ses concepts sont définis de manière non ambiguë par des axiomes (cf. 2.2.2). Par exemple, la *Systematized Nomenclature of Medicine Clinical Terms* (**SNOMED CT**) est une ontologie décrivant de manière formelle les relations hiérarchiques et transversales entre les concepts médicaux, définis de façon explicite.

La table 2.1 synthétise les représentations de connaissances mentionnées dans cette thèse.

Nous exprimons deux limites à l'utilisation des terminologies pour représenter des connaissances extraites à partir d'entreposés de données :

- Les terminologies sont des modèles de données exclusivement théoriques. Un modèle de données théorique [21] est une représentation abstraite et structurée des concepts, des entités et des relations d'un domaine particulier, sans inclure de données spécifiques ou instances concrètes. Il sert de plan ou de schéma pour organiser et comprendre les informations dans ce domaine. Les modèles théoriques définissent les classes, les propriétés et les relations entre les concepts, mais ils ne contiennent pas les valeurs spécifiques ou les données réelles associées à ces concepts. Ils offrent une abstraction, une structuration et une standardisation des concepts d'un domaine, mais ne sont pas instanciables.
- Les terminologies n'ont pas de sémantique formelle associée, et pour ainsi elles ne permettent pas de faire de raisonnement automatique.

Les ontologies sont quant à elles, à la fois des modèles de données théoriques et instances. Elles permettent non seulement de représenter un domaine, mais aussi d'intégrer des données concrètes (*i.e.*, des entités) au modèle représenté. Les modèles théoriques créés sont voués à être instanciés dans des cas pratiques dans le but de concrétiser les données d'application, les stocker, les rendre manipulables et requêtables.

En ce sens, les ontologies sont divisées en deux composants principaux :

- La **TBox** (*Terminological Box*) contient les axiomes terminologiques, c'est-à-dire les déclarations générales sur la structure du domaine, comme les définitions de concepts et les relations entre eux.
- L'**ABox** (*Assertional Box*) contient les assertions, c'est-à-dire les faits concrets sur des instances spécifiques dans le domaine.

Par exemple, la TBox peut définir que tous les patients sont des personnes, tandis que l'ABox peut contenir l'information que "Bob est un patient" et que "Bob a une œsophagite peptique".

Un formalisme particulier qui permet de construire des ontologies est celui proposé par les logiques de descriptions.

Modèle de données	Structure	Brève description
Common Terminology Criteria for Adverse Events (CTCAE) [31]	Classification	Répertorie les effets indésirables connus de traitements et les classes en cinq grades de sévérité, selon des critères précis.
World Health Organization Adverse Reactions Terminology (WHOART) [216]	Classification	Anciennement utilisée pour la codification des effets indésirables des médicaments dans les essais cliniques et la pharmacovigilance, mais tend à être remplacée par MedDRA.
Medical Dictionary for Regulatory Activities (MedDRA) [129]	Thésaurus	Vocabulaire médical utilisé pour coder les informations sur les médicaments, les effets indésirables et les symptômes, assurant une communication uniforme dans les essais cliniques et la pharmacovigilance.
Anatomical Therapeutic Chemical Classification System (ATC) [219]	Classification	Système de classification des médicaments basé sur les organes ou systèmes sur lesquels ils agissent et leurs propriétés thérapeutiques.
International Classification of Diseases (ICD-11) [220]	Classification	Système de classification des maladies et des problèmes de santé, utilisé pour coder les diagnostics et les causes de décès.
Unified Medical Language System (UMLS) [206]	Métathésaurus	Répertoire de termes interconnectés par des relations hiérarchiques, synonymiques, et associatives, provenant de nombreuses terminologies.
Medical Subject Headings (MeSH) [132]	Thésaurus	Vocabulaire contrôlé utilisé pour indexer les articles dans PubMed et d'autres bases de données de la National Library of Medicine (NLM).
National Cancer Institute Thesaurus (NCIt) [139]	Thésaurus	Vocabulaire contrôlé et hiérarchisé utilisé pour décrire les concepts en oncologie, facilitant la recherche et l'échange de données.
Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [189]	Ontologie	Système de terminologie clinique complet qui fournit des codes, des termes, des synonymes et des définitions pour documenter les soins de santé.
Logical Observation Identifiers Names and Codes (LOINC) [119]	Ontologie	Norme pour identifier les observations médicales de manière unique, incluant des tests de laboratoire, des mesures cliniques et des documents.
HemOnc.org (HemOnc) [87] [213] [212] [173]	Ontologie	Ontologie spécialisée dans les traitements oncologiques et hématologiques, notamment les schémas thérapeutiques de chimiothérapies.
Time Event Ontology (TEO) [113] [35]	Ontologie	Ontologie centrée sur la représentation des événements médicaux et leurs relations temporelles.
Cancer Care Treatment Outcome Ontology (CCTOO) [115]	Ontologie	Ontologie dédiée à la représentation des réponses aux traitements des cancers.
Drug Ontology (DRON) [54]	Ontologie	Ontologie dédiée à la représentation des connaissances sur les médicaments, leurs ingrédients et leurs différentes catégories.
Ontology for Biomedical Investigations (OBI) [150]	Ontologie	Ontologie destinée à représenter les investigations biomédicales, incluant des concepts pour les protocoles expérimentaux et les matériaux.
RxNorm [174]	Thésaurus	Norme de nomenclature pour les médicaments cliniques, facilitant l'interopérabilité entre différents systèmes de prescription et de pharmacie.
Chemical Entities of Biological Interest (ChEBI) [34]	Ontologie	Ontologie de composés chimiques d'intérêt biologique, contenant des informations sur leurs actions chimiques et leurs rôles biologiques.
Référentiel Ouvert du Médicament (Romedi) [171] [42]	Ontologie	Ontologie qui lie les médicaments et leurs noms français à des ressources médicamenteuses internationales (ATC, DrugBank).
DrugBank (DrugBank) [55]	Ontologie	Ontologie sur les médicaments et les cibles médicamenteuses, incluant des informations sur la chimie, la pharmacologie et les effets secondaires.
Pharmacogenomics Linked Open Data (PGxLOD) [158] [135] [134]	Ontologie	Graphe de connaissances intégrant des unités pharmacogénomiques, (association gène-médicament-phénotype) produites à partir de diverses sources publiques, de la littérature et cliniques
FORUM [49]	Ontologie	Graphe de connaissances intégrant des données de diverses sources publiques et de la littérature scientifique pour inférer des associations entre produits chimiques et maladies.

TABLE 2.1 – Représentations de connaissances bio-médicales mentionnées dans cette thèse.

2.2.2 Bases des logiques de descriptions

Les logiques de description (*Description Logics*, DL) sont une famille de langages de représentation de connaissances utilisées dans les ontologies pour formaliser la connaissance d'un domaine en décrivant ses concepts et leur sémantique [103].

Une ontologie \mathcal{O} est définie par un ensemble d'axiomes logiques Φ , construits à partir d'un ensemble d'individus $\mathcal{I} = \{i, j, \dots\}$, de concepts $\mathcal{C} = \{C, D, \dots\}$, de rôles $\mathcal{R} = \{R, S, \dots\}$ et de constructeurs \mathcal{S} qui combinent ces concepts et rôles.

Définition - Individus

Les individus sont les entités spécifiques du domaine représenté par l'ontologie. Ce sont les instances concrètes des concepts.

Par exemple, dans une ontologie médicale, un individu pourrait être un patient spécifique, tel que "Bob".

Définition - Concepts

Les concepts sont des catégories ou classes qui regroupent des individus partageant des caractéristiques communes.

Par exemple, le concept de "Patient" regroupe tous les individus qui sont des patients.

Définition - Rôles

Les rôles sont des relations entre les individus. Ils définissent comment les individus interagissent entre eux ou avec des valeurs de données.

Par exemple, le rôle "a un diagnostic de" pourrait relier un patient à une maladie telle que l'œsophagite peptique.

Définition - Constructeurs

Les constructeurs sont des opérateurs utilisés pour créer des concepts et des rôles plus complexes à partir de concepts et de rôles plus simples. Ils permettent de définir des relations et des contraintes entre les concepts.

Par exemple, l'intersection (\sqcap) combine deux concepts pour créer un concept qui inclut tous les individus appartenant aux deux concepts initiaux.

Remarque - principaux types d'axiomes dans les ontologies

Les principaux types d'axiomes dans les ontologies incluent :

- $C \sqsubseteq D$ (subsumption) : Un concept C est un sous-ensemble d'un autre concept D .
- $C \equiv D$ (équivalence) : Deux concepts C et D sont équivalents.

- $C(a)$ (instantiation) : Un individu a est une instance du concept C .
- $R(i, j)$ (relation) : Une relation R existe entre les individus i et j .
- $R^{-1}(j, i)$ (relation inverse) : Si une relation R existe entre les individus i et j , alors une relation inverse R^{-1} existe entre les individus j et i .

Par exemple, un axiome de subsomption pourrait indiquer que tous les patients atteints d'œsophagite peptique sont des patients.

2.2.3 Le raisonnement automatique

Le raisonnement automatique utilise les axiomes et les définitions de l'ontologie pour inférer des relations telles que la subsomption, la transitivité et les relations inverses. Ce type de raisonnement repose sur les logiques de description et permet de déduire automatiquement de nouvelles connaissances à partir de la structure existante de l'ontologie.

Par exemple, considérons une ontologie où nous avons les concepts suivants :

- $C \sqsubseteq D$: Tous les patients atteints d'œsophagite peptique (C) sont des patients atteints d'œsophagite (D).
- $D \sqsubseteq E$: Tous les patients atteints d'œsophagite (D) sont des patients ayant une maladie digestive (E).

En utilisant le raisonnement automatique, nous pouvons inférer la nouvelle relation suivante :

- Par transitivité : $C \sqsubseteq E$: Tous les patients atteints d'œsophagite peptique (C) sont des patients ayant une maladie digestive (E).

2.2.4 Ontologies, bases de connaissances, système à base de connaissances, graphes de connaissances

Les notions d'"ontologie", "base de connaissances", "graphe de connaissances" et "système à base de connaissances" sont liées, et leurs frontières sont difficiles à établir. Dans le cadre de cette thèse, nous posons les définition de ces notions en nous appuyant notamment sur celles de BAADER [11] pour les distinguer.

Les bases de connaissances, graphes de connaissances et systèmes à base de connaissances intègrent généralement une/des ontologie(s). La différence des qualificatifs employés pour ces quatre notions réside surtout dans l'/les **usage(s)** de l'ontologie ou du réseau d'ontologies. Pour clarifier cela, on commence donc par redéfinir la notion d'ontologie, de manière plus détaillée.

Lorsque l'ontologie est instanciée et que l'on met à disposition un moteur de recherche et de requête en ligne pour l'utilisateur, bien souvent on préfère parler de base de connaissances.

Base de connaissances Une base de connaissances est une instantiation d'une ontologie qui inclut des individus, ou instances, organisés selon les concepts et les relations définies dans l'ontologie. Une base de connaissances contient une boîte terminologique (TBox) avec des axiomes relatifs aux concepts et aux relations, et une boîte assertionnelle (ABox) avec

des faits sur les individus, exprimant les relations entre eux et indiquant quelles concepts ils instancient. La base de connaissances permet non seulement de stocker des informations, mais aussi de les interroger et de les analyser pour en extraire des connaissances supplémentaires.

Par exemple, dans la table 2.1, on utilise plutôt le terme de base de connaissances, voire de base de données pour la qualifier la [DrugBank](#) ou [Romedi](#).

Lorsque l'on souhaite insister sur le développement d'un réseau de bases de connaissances, on emploie le terme de graphe de connaissance.

Graphe de connaissances Il n'existe pas de définition formelle et unanimement acceptée au sein de la communauté pour les graphes de connaissances [60, 89, 65], mais plusieurs propriétés essentielles sont généralement reconnues. Premièrement, un graphe de connaissances est un type de graphe : il modélise des entités sous forme de noeuds et les relations entre ces entités sous forme d'arcs. En tant que graphe de connaissances, il décrit des entités d'un domaine à l'aide de concepts et d'instances, avec des relations qui sont dirigées et étiquetées pour représenter leurs connexions. Cela le rend similaire aux ontologies et aux bases de connaissances. Toutefois, trois distinctions principales sont admises :

- Un graphe de connaissances est généralement de grande taille, et comporte un grand nombre de noeuds et de liens.
- Un graphe de connaissances peut intégrer plusieurs ontologies, et couvre ainsi divers domaines.
- Un graphe de connaissances met l'accent sur les instances des concepts définis par les ontologies, et enrichit ainsi les descriptions des entités et leurs interrelations.

Par exemple, dans la table 2.1, [PGxLOD](#) [135], et [FORUM](#) [49] sont des exemples de graphes de connaissances. Ils respectent les critères submentionnés des graphes de connaissances par leurs grands nombres de noeuds et d'arcs, leur intégration de plusieurs ontologies couvrant divers domaines, et leur accent sur les instances des concepts définis par les ontologies.

La capacité de raisonnement associées aux ontologies, bases et graphes de connaissances permet le développement d'outils d'aide à la décision avec les systèmes à base de connaissances.

Système à base de connaissances Un système à base de connaissances est une application informatique qui utilise une base de connaissances pour résoudre des problèmes complexes, prendre des décisions, ou fournir des recommandations [1]. Ces systèmes s'appuient sur des techniques de raisonnement logique pour interpréter et manipuler les connaissances stockées, et pour générer de nouvelles connaissances ou actions en réponse à des requêtes spécifiques. L'ensemble de la base de connaissances et le moteur de raisonnement constituent le système à base de connaissances.

Par exemple, ACHOUR et al. [2] ont développé une base de connaissance et un corpus de règles basés à la fois sur l'[UMLS](#) et l'expertise médicale pour fournir des recommandations sur les transfusions sanguines.

Toutefois, l'utilisation d'outils de raisonnement logique est souvent limitée par la taille de l'ontologie ou du graphe de connaissances. Pour explorer les graphes de connaissances

de grande taille, on emploie donc souvent partiellement ces raisonnements avec des règles simples, et on les combine à des méthodes d'exploration de graphes.

Méthodes d'exploration des graphes de connaissances Les méthodes d'exploration des graphes de connaissances visent à analyser et extraire des informations pertinentes à partir de graphes de connaissances de grande taille. Elles comprennent des techniques telles que les plongements de graphes (*graph embedding*) et les alignements d'ontologies (*ontology matching*). Ces méthodes permettent de découvrir des relations, d'aligner des entités similaires entre différentes ontologies et d'améliorer la qualité des connaissances intégrées. Par exemple MONNIN et al. [134] ont utilisé des réseaux de neurones convolutifs sur graphes (*Graph Convolutional Networks*, GCN) pour découvrir des alignements entre des entités de PGxLOD, facilitant ainsi l'identification et l'appariement automatique d'entités similaires issues de différentes sources de données pharmacogénomiques.

2.3 Représenter les connaissances avec les outils du Web sémantique

Le Web sémantique, introduit par Tim Berners-Lee en 2001 [19], vise à enrichir le Web en y ajoutant des significations sémantiques pour permettre une meilleure compréhension du contenu du Web par les machines. Le Web sémantique apporte des technologies standardisées pour la représentation des connaissances et l'implémentation des ontologies. Dans cette section, nous nous intéressons particulièrement aux standards et technologies du Web sémantique qui facilitent la création et la gestion d'ontologies.

2.3.1 Les outils du Web sémantique pour implémenter la représentation d'un domaine dans la pratique

Les technologies du Web Sémantique offrent des outils permettant de modéliser les ontologies, en fournissant des moyens de représenter les notions essentielles telles que les individus, les concepts, et d'autres éléments décrits dans la section 2.2.2. La nomenclature reprend celle de la programmation orientée objet (classe pour concept et propriété pour rôle). Nous décrivons les principales technologies avec des exemples tirés de la SNOMED CT. Au cours de cette thèse, les deux ontologies qui ont été développées (chapitre 4) ont été conçues à l'aide d'un module python, Owlready, développé par Jean-Baptiste Lamy. Les différentes définitions apportées dans cette section s'appuient sur son livre "Python et les ontologies" [103] et sur les recommandations du World Wide Web Consortium (W3C), OWL [209] et RDFS [210].

2.3.1.1 Les classes pour représenter des concepts

Les classes constituent les éléments de base des ontologies. Elles représentent des concepts abstraits qui regroupent des entités partageant des caractéristiques communes. Par exemple, dans la SNOMED CT, des classes comme "Disease" (maladie), "Procedure" (procédure) ou "Clinical Finding" (observation clinique) sont utilisées pour structurer les connaissances médicales. Ces classes permettent de modéliser les différents éléments d'un domaine en définissant clairement les concepts et leurs attributs.

2.3.1.2 La propriété "rdf :type" pour les instances (ou individus)

Les instances, ou individus, sont les entités concrètes d'un cas d'application de l'ontologie qui instancient les classes définies dans l'ontologie. Chaque instance représente un élément spécifique et identifiable du cas d'application. Pour indiquer qu'une entité est une instance d'une classe, on utilise généralement la propriété **rdf : type**. Cette propriété relie une instance à sa classe, exprimant ainsi que cette instance appartient à une catégorie spécifique définie par la classe.

Par exemple, Bob, identifié par son numéro patient et atteint d'une œsophagite peptique instancie le concept "Patient" en utilisant la propriété **rdf : type**.

2.3.1.3 Les propriétés pour les relations hiérarchiques et transversales

Les relations dans les ontologies peuvent être hiérarchiques ou transversales, ce qui permet de structurer les connaissances de manière flexible et détaillée.

La propriété "rdfs :subClassOf" pour les relations hiérarchiques Aussi appelées relations de subsomption, les relations hiérarchiques permettent de créer des taxonomies de concepts. Elles relient deux classes de l'ontologie, à l'aide de la propriété de subsomption **rdfs : subClassOf**.

Par exemple, : Dans la **SNOMED CT**, la classe **16761005:Oesophagitis (disorder)** est une sous-classe de **"373407002:Inflammatory disorder of digestive system (disorder)"**, ce qui signifie que toutes les instances de **16761005:Oesophagitis (disorder)** sont également des instances de **"373407002:Inflammatory disorder of digestive system (disorder)"**.

Les propriétés pour représenter des relations transversales Les propriétés transversales permettent de créer des relations horizontales, non hiérarchiques. Contrairement aux relations hiérarchiques, il existe de multiples propriétés transversales, et elles peuvent être définies par l'utilisateur au moment de la conception de l'ontologie. Ces propriétés permettent une grande liberté d'expressivité. Il en existe deux types :

- **Les propriétés d'objet** : Elles relient des instances de deux classes différentes.
Par exemple, : La propriété d'objet **363698007:Finding site** peut relier des instances de la classe **16761005:Oesophagitis (disorder)** aux instances de la classe **32849002:Eso-phageal structure**.
- **Les propriétés de données** : Elles relient des instances de classe à des valeurs littérales. Une valeur littérale est une donnée concrète, comme un nombre ou une chaîne de caractères. Par exemple, la propriété "has_age" peut relier une instance de la classe "Patient" à une valeur littérale représentant l'âge du patient.

2.3.2 Resource Description Framework (RDF)

Le Resource Description Framework (RDF) représente les informations sous forme de triplets, composés d'un sujet, d'un prédicat et d'un objet, qui peuvent être visualisés comme des noeuds et des arcs dirigés et étiquetés dans un graphe.

Un triplet RDF se compose de :

- **Sujet** : La ressource ou l'entité sur laquelle porte la déclaration.
- **Prédicat** : La propriété ou le type de relation entre le sujet et l'objet.
- **Objet** : La valeur ou l'autre ressource reliée au sujet par le prédicat.

Un exemple de triplet dans le contexte de la **SNOMED CT** pourrait être :

```
(sujet: http://example.org/instance/oesophagitis123,
prédicat: http://www.w3.org/1999/02/22-rdf-syntax-ns#type,
objet: http://snomed.info/id/16761005)
(sujet: http://example.org/instance/oesophagitis123,
prédicat: http://snomed.info/field/findingSite,
objet: http://example.org/instance/esophagus456)
(sujet: http://example.org/instance/esophagus456,
prédicat: http://www.w3.org/1999/02/22-rdf-syntax-ns#type,
objet: http://snomed.info/id/32849002)
(sujet: http://snomed.info/id/16761005,
prédicat: http://www.w3.org/2000/01/rdf-schema#subClassOf,
objet: http://snomed.info/id/373407002)
```

Le premier triplet décrit que l'instance spécifique "oesophagitis123" est une instance de la classe "16761005 :Oesophagitis (disorder)", le second triplet décrit que "oesophagitis123" a pour site anatomique une instance spécifique de l'œsophage "esophagus456", le troisième triplet décrit que "esophagus456" est une instance de la classe "32849002 :Esophageal structure", et le quatrième triplet décrit que la classe "16761005 :Oesophagitis (disorder)" est une sous-classe de "373407002 :Inflammatory disorder of digestive system (disorder)".

RDF offre une flexibilité et une applicabilité universelle en permettant de représenter n'importe quelle entité, abstraite ou physique, sur le Web. Les triplets RDF permettent de décrire les relations entre ces entités, facilitant ainsi la modélisation des connaissances et l'intégration de données hétérogènes provenant de diverses sources.

2.3.3 Identifier les ressources sur le Web avec les URI

Pour identifier de manière unique les ressources sur le Web, le Web sémantique utilise des URI (Uniform Resource Identifiers). Un URI est semblable à une URL (Uniform Resource Locator), mais l'URI est défini plus largement. Un URI identifie une ressource sur le Web. Par exemple, dans un triplet RDF ci-dessous :

```
(sujet: http://example.org/instance/oesophagitis123,
prédicat: http://www.w3.org/1999/02/22-rdf-syntax-ns#type,
objet: http://snomed.info/id/16761005)
```

- `http://example.org/instance/oesophagitis123` : Ici, "`http://example.org/`" est un espace de nommage (namespace) que nous avons créé pour notre exemple. Il identifie le domaine de l'ontologie créée, et "`instance/oesophagitis123`" identifie une instance spécifique d'œsophagite dans notre cas d'application.
- `http://www.w3.org/1999/02/22-rdf-syntax-ns#type` : C'est l'URI du prédicat "type" dans la syntaxe RDF, définissant que le sujet est une instance de la classe spécifiée par l'objet. L'URI ici est standardisé et fournit un moyen unique d'identifier cette propriété RDF.
- `http://snomed.info/id/16761005` : C'est l'URI identifiant la classe "oesophagite" dans la SNOMED CT, une ressource accessible sur le Web. Cette URI permet de référencer de manière unique cette classe spécifique dans une ontologie reconnue.

Les URI permettent ainsi de décrire les relations entre les entités de manière unique et interopérable.

2.3.4 Des modèles de référence à réutiliser

Il est essentiel d'utiliser des modèles de référence pour représenter certains domaines ou concepts, en particulier lorsqu'ils disposent de représentations standardisées et acceptées par la communauté [73]. Ces modèles offrent une normalisation et une interopérabilité qui sont cruciales pour l'intégration et l'échange de données entre différentes institutions et systèmes. Dans le cadre de cette thèse, axée sur la représentation des chimiothérapies et de leurs réponses, nous mettons en avant certains modèles de référence particulièrement pertinents dans ce domaine (*cf.* section 4.1.2 du chapitre 4).

2.3.4.1 Représenter le temps avec la Time Ontologie

La Time Ontology (TO) [45], proposée par le World Wide Web Consortium (W3C), est une ontologie standardisée pour représenter les concepts temporels. Elle permet de décrire des points dans le temps, des intervalles, des durées et des relations temporelles. Toute entité d'une ontologie peut être le sujet de la propriété `time:hasTime`, la liant à une entité de la TO (`time:TemporalEntity`) telle `time:Instant` ou `time:ProperInterval`. Les principaux éléments de la Time Ontology incluent des classes telles que `time:Instant`, représentant un point unique dans le temps, et `time:ProperInterval`, représentant une durée de temps délimitée par un début et une fin. Des propriétés comme `time:inXSDDateTime`, qui associe un instant à une date et une heure précises, et `time:hasBeginning` et `time:hasEnd`, qui définissent respectivement le début et la fin d'un intervalle, sont essentielles pour structurer les données temporelles. La propriété `time:hasXSDDuration` permet de spécifier la durée d'un intervalle.

2.3.4.2 Représenter la provenance des connaissances avec PROV-O

PROV-O est une ontologie développée par le W3C pour représenter les informations de provenance, c'est-à-dire l'origine des données et des connaissances dans des graphes de connaissances. PROV-O utilise des concepts clés tels que `PROVO:Entity`, représentant des objets physiques, numériques ou abstraits, `PROVO:Activity`, qui désigne des actions ou processus produisant des entités, et `PROVO:Agent`, qui sont les acteurs responsables des activités, comme des personnes, des organisations ou des logiciels.

Par exemple, dans le cadre de PGxLOD [135], PROV-O est utilisée pour détailler la provenance de chaque relation ternaire de pharmacogénomique (association gène-médicament-effet indésirable). Ceci est particulièrement pertinent car ces relations proviennent de diverses sources telles que PharmGKB, PubMed et les dossiers patients de l'HEGP. Les autorités qui publient les sources à partir desquelles les unités de connaissances sont extraites sont considérées comme des agents (p. ex. l'équipe de PharmGKB, la National Library of Medicine en charge de PubMed, ou un hôpital en charge d'un entrepôt de dossiers médicaux électroniques). Les sources de données (p. ex. une version de PharmGKB, de PubMed, ou un entrepôt de dossiers médicaux électroniques) sont attribuées à des agents et peuvent ensuite être utilisées pour dériver des données. Ces données, à leur tour, sont utilisées pendant l'exécution d'une activité (p. ex. un algorithme de fouille de données). L'utilisation de PROV-O assure ainsi la transparence des données intégrées, facilitant la traçabilité et l'audit des informations issues de sources multiples.

2.3.4.3 Dans le domaine biomédical

L'UMLS carrefour des terminologies médicales L'*Unified Medical Language System* (UMLS) est une ressource développée par la National Library of Medicine (NLM) qui intègre et normalise de nombreuses terminologies médicales. L'UMLS permet de relier des concepts issus de différentes terminologies grâce à une hiérarchisation rigoureuse et des relations sémantiques bien définies. Parmi les nombreuses terminologies liées à l'UMLS, on trouve notamment les terminologies de toxicités MedDRA et WHOART. Cette structuration hiérarchique est particulièrement précieuse pour le raisonnement automatique, car elle permet d'inférer de nouvelles connaissances à partir des relations existantes.

Par exemple, l'arborescence des concepts dans l'UMLS peut être exploitée pour développer des systèmes à base de connaissances, comme illustré par le système de support aux décisions cliniques pour les transfusions sanguines décrit par ACHOUR et al. [2]. L'UMLS offre également des opportunités pour la normalisation et l'interopérabilité entre divers systèmes de santé, en facilitant l'intégration et l'analyse des données de différentes sources.

MedDRA-WHOART et le CTCAE pour les toxicités oncologiques et leurs sévérités Plusieurs terminologies sont disponibles pour la classification des effets indésirables. Le *Medical Dictionary for Regulatory Activities* (MedDRA) est un thésaurus bio-pharmaceutique de référence qui offre un répertoire de termes codifiés concernant les maladies, les symptômes et les effets indésirables, et est largement utilisé lors des essais cliniques pour standardiser la communication des événements indésirables. Avant l'adoption de MedDRA, le *World Health Organization Adverse Reactions Terminology* (WHOART) était utilisé pour coder spécifiquement les effets indésirables des médicaments. Comme discuté dans les sections 1.2.1.2 et 1.2.3 du chapitre 1, le *Common Terminology Criteria for Adverse Events* (CTCAE) est couramment utilisé au cours des chimiothérapies pour évaluer la sévérité des toxicités observées chez les patients. Les toxicités du CTCAE sont alignées avec MedDRA, ce qui facilite cette intégration. Dans le cadre de la détection et de l'intégration des toxicités dans un modèle de connaissances, il est pertinent de tirer parti de ces trois terminologies et de leurs interconnexions avec l'UMLS. Utiliser MedDRA, WHOART et CTCAE ensemble permet de construire un dictionnaire exhaustif et standardisé des toxicités chimio-induites. (cf. sections 3.1.1.1 du chapitre 3 et 4.2.2.1 du chapitre 4)

L'ATC pour les médicaments L'*Anatomical Therapeutic Chemical (ATC) Classification System* est un système de classification des médicaments développé par l'OMS. Il classe les médicaments en fonction de l'organe ou du système sur lequel ils agissent et de leurs propriétés thérapeutiques. Chaque médicament reçoit un code unique qui permet de l'identifier et de le classer de manière standardisée. L'ATC est largement utilisé dans les études pharmaco-épidémiologiques et les systèmes de gestion de la santé pour standardiser la classification et l'analyse des données sur les médicaments. Il facilite également la comparaison des données sur les médicaments entre différents pays et institutions, améliorant ainsi la cohérence et l'interopérabilité des informations sur les traitements médicaux.

2.3.5 Gérer et requêter un graphe de connaissances avec un triple store et SPARQL

Système de gestion de bases de connaissances

Les bases de connaissances peuvent être vues comme des ensembles de fichiers contenant des triplets RDF. Ces triplets peuvent être stockés et interprétés dans des bases de données graphes, appelées des triplestores. De nombreux triplestores existent tel que GraphDB [154] ou Virtuoso [190]. Ils s'utilisent via une interface graphique ou en API. De la même façon que les systèmes de bases de données, ils permettent d'administrer la base, de définir des droits à des utilisateurs, définir des schémas, supprimer ou ajouter des triplets, et surtout d'interroger les données avec le langage de requête SPARQL.

Le langage SPARQL

SPARQL est un langage de requête et un protocole conçu pour interroger et manipuler les bases de connaissances. En SPARQL, l'interrogation de ces bases de connaissances se fait en recherchant des correspondances entre des motifs décrits avec des triplets dans la requête et les triplets RDF instanciés présents dans la base. SPARQL utilise des motifs de graphes, qui sont des ensembles de triplets pouvant inclure des variables représentant des sujets, des prédicts ou des objets. Une requête SPARQL recherche des correspondances entre ces motifs et les triplets du graphe RDF.

Dans la plupart des cas, une requête SPARQL inclut une clause SELECT qui permet de sélectionner les variables à inclure dans le résultat, une clause FROM pour définir le graphe à interroger et une clause WHERE qui contient les motifs de graphes. SPARQL permet également d'effectuer des opérations telles que des unions, des filtres ou des motifs optionnels.

En plus de la clause SELECT, SPARQL propose d'autres types de requêtes : la clause CONSTRUCT permet de construire un nouveau graphe en fonction des résultats de la requête, la clause DESCRIBE fournit une description d'une ressource spécifique, et la clause ASK vérifie l'existence d'un motif précis dans le graphe RDF.

Le requêtage en SPARQL d'une base de connaissances peut être rendu accessible sur le web via un point d'entrée SPARQL (SPARQL endpoint). Par exemple le point d'entrée de PGxLOD est accessible sur <https://pgxlod.loria.fr/sparql> et celui de FORUM sur <https://forum.semantic-metabolomics.fr/sparql/>. La requête SPARQL suivante :

```
SELECT COUNT(DISTINCT ?entity)
WHERE
{
    ?entity a <http://pgxo.loria.fr/PharmacogenomicRelationship>
}
```

permet par exemple de savoir que 50 435 unités pharmacogénomiques ont été instanciées dans le graphe de PGxLOD.

Analyse de données

En plus de permettre la navigation au sein du graphe, les requêtes SPARQL permettent d'analyser quantitativement le graphe. La page [Statistics](#) de PGxLOD offre un aperçu des

résultats de requêtes qui permettent par exemple de connaître la répartition des sources qui ont généré des unités pharmacogénomiques. Le graphe de connaissances de FORUM a en partie été enrichi avec de l'analyses de données avec SPARQL. En effet, pour établir un lien entre un composé chimique et son concept biomédical, la significativité du lien a été évaluée avec des résultats de requêtes SPARQL [49].

2.4 Raisonnement et fouille de graphe

2.4.1 Le raisonnement dans un système à base de connaissances

Le raisonnement dans un système à base de connaissances peut être principalement classé en deux catégories : le raisonnement automatique et le raisonnement basé sur des règles. Le raisonnement automatique repose sur les logiques de description pour inférer des connaissances dans le système.

Par exemple GALOPIN et al. [68] ont développé un système à base de connaissances pour la gestion des patients atteints de multiples troubles chroniques. Ce système utilise le raisonnement sémantique pour enrichir les descriptions des patients en ajoutant des concepts dérivés. Par exemple, un patient souffrant d'hypertension et de diabète de type 2 est automatiquement associé à des concepts comme "Maladie cardiovasculaire" et "Maladie vasculaire". En utilisant la hiérarchie de subsumption des profils de patients, le système détecte et résout les conflits entre les recommandations des différentes lignes directrices cliniques. Ainsi, il évite de recommander des médicaments inappropriés pour les patients avec des comorbidités, en proposant des alternatives plus sûres et spécifiques.

Le raisonnement basé sur des règles utilise des règles logiques supplémentaires, comme celles définies en Semantic Web Rule Language (SWRL), pour inférer de nouvelles connaissances qui ne sont pas directement déduites des axiomes de l'ontologie. SWRL permet de définir des règles conditionnelles qui peuvent combiner plusieurs concepts et relations pour inférer des conclusions plus complexes.

Par exemple CHANDRA et al. [32] utilisent des règles SWRL pour inférer des recommandations de traitement pour les patients diabétiques. Ce système utilise des règles spécifiques pour analyser les données des patients et recommander des médicaments anti-diabétiques appropriés en fonction de leurs conditions médicales et de leurs antécédents.

Dans cette thèse, nous nous sommes intéressés au raisonnement temporel. Le raisonnement temporel permet d'inférer des relations temporelles. Il peut se baser sur l'algèbre d'Allen [4]. Celle-ci définit 13 relations qui couvrent toutes les interactions possibles entre deux intervalles temporels. Ces relations sont exposées dans la table 2.2.

Dans les ontologies décrivant des intervalles temporels, ces relations peuvent être inférées à l'aide de règles SWRL. BATSAKIS et al. [15] démontrent qu'il est possible d'inférer les relations d'Allen en utilisant deux ontologies spécifiques : la Time Ontology (*cf.* 2.3.4.1) et la Time Event Ontology [113]. Ils ont intégré des règles SWRL dans des systèmes à base de connaissances pour déduire les relations de Allen entre des intervalles temporels. BATSAKIS et al. [15] avertissent cependant de la complexité exponentielle d'un raisonnement inférant toutes les relations de Allen. Les inférences de raisonnement temporel peuvent être transformées en un problème de satisfaction de contraintes (CSP). Lorsqu'une relation qualitative

Relation	Relation inverse	Schéma	Séquence chronologique
X before Y	Y after X	$\underline{\quad X \quad}$ $\underline{\quad Y \quad}$	$X_{start} < X_{end} < Y_{start} < Y_{end}$
X equals Y	Y equals X	$\underline{\quad X \quad}$ $\underline{\quad Y \quad}$	$X_{start} = Y_{start} \quad X_{end} = Y_{end}$
X meets Y	Y met by X	$\underline{\quad X \quad}$ $\underline{\quad Y \quad}$	$X_{start} < X_{end} = Y_{start} < Y_{end}$
X overlaps Y	Y overlapped by X	$\underline{\quad X \quad}$ $\underline{\quad Y \quad}$	$X_{start} < Y_{start} < X_{end} < Y_{end}$
X contains Y	Y during X	$\underline{\quad X \quad}$ $\underline{\quad Y \quad}$	$X_{start} < Y_{start} < Y_{end} < X_{end}$
X starts Y	Y started by X	$\underline{\quad X \quad}$ $\underline{\quad Y \quad}$	$X_{start} = Y_{start} < X_{end} < Y_{end}$
X finishes Y	Y finished by X	$\underline{\quad X \quad}$ $\underline{\quad Y \quad}$	$Y_{start} < X_{start} < Y_{end} = X_{end}$

TABLE 2.2 – Algèbre d’Allen. X et Y décrivent deux intervalles temporels, délimités par leurs instants de début X_{start} , Y_{start} et leurs instants de fins X_{end} , Y_{end}

est établie entre deux entités temporelles, cela impose des contraintes sur les relations que ces entités peuvent avoir avec d’autres entités temporelles. Par exemple, si l’événement A se produit avant l’événement B, alors A doit nécessairement être avant tout autre événement qui suit B. Ces contraintes restrictives créent un réseau complexe de dépendances à vérifier. Ce CSP est connu pour être NP-complets, ce qui signifie qu’il n’existe pas de solution efficace en temps polynomial dans le cas général. Cette complexité s’applique particulièrement lorsque les relations temporelles qualitatives doivent être évaluées de manière cohérente à travers de nombreuses entités temporelles.

Cette complexité computationnelle liée au raisonnement automatique n’est pas seulement limitée au raisonnement temporel.

2.4.1.1 Les raisonneurs et leurs limites computationnelles

Les raisonneurs, tels que HermiT [136] et Pellet [187], sont des outils utilisés pour effectuer du raisonnement. Cependant, bien que ces raisonneurs soient puissants, ils sont confrontés à des limites computationnelles. Le raisonnement automatique, en particulier lorsqu’il s’agit de relations qualitatives complexes ou d’ontologies de grande taille, peut devenir un problème NP-difficile, ce qui rend le temps de calcul nécessaire exponentiellement long avec l’augmentation de la taille et de la complexité des données. Cette complexité est due à la nature des logiques description et des problèmes de satisfaction de contraintes,

qui nécessitent des ressources computationnelles importantes pour assurer la cohérence et l'inférence des nouvelles connaissances.

Ainsi, bien que les raisonneurs soient utiles dans des systèmes basés sur les connaissances, leurs limitations en termes de performance imposent des défis significatifs dans l'analyse des graphes de connaissances.

2.4.2 Vers la fouille des graphes de connaissances

La fouille des graphes de connaissances implique l'utilisation de diverses techniques pour extraire des informations utiles et des relations cachées au sein des graphes. Ces techniques permettent de naviguer à travers les structures complexes de ces graphes, de découvrir des motifs récurrents, et d'établir des liens significatifs entre les différentes entités et relations. Parmi les méthodes les plus couramment utilisées, on trouve :

- L'alignement ontologique : Il s'agit de mettre en correspondance les concepts et les relations de différentes ontologies pour intégrer les données provenant de sources hétérogènes.
- La recherche de sous-graphes fréquents : Cette technique consiste à identifier des motifs fréquents, des sous-graphes intéressants, et des communautés au sein des graphes. Elle est utilisée pour découvrir des modèles cachés, des anomalies et des structures sous-jacentes.
- La prédiction de lien : Cette tâche consiste à prédire les relations manquantes ou futures entre les entités d'un graphe, en se basant sur les motifs observés et les structures existantes.
- Le clustering : Cette technique vise à regrouper les noeuds ou les sous-graphes similaires en clusters, facilitant ainsi l'identification de communautés et de structures significatives.
- Les plongements de graphes (*graph embeddings*) : Les plongements de graphes sont des méthodes qui visent à représenter les entités et les relations d'un graphe de connaissances dans un espace vectoriel de faible dimension tout en préservant la structure du graphe. Ces représentations vectorielles permettent d'appliquer des techniques d'apprentissage automatique pour des tâches telles que la classification de noeuds, la prédiction de liens, le clustering, l'alignement ontologique, et la recherche de sous-graphes fréquents. Les plongements de graphes sont souvent utilisés comme étape préalable pour exécuter ces autres tâches.

Dans le cas de la fouille de graphes de connaissances bio-médicaux, on combine souvent ces techniques avec du raisonnement automatique pour enrichir les résultats et améliorer la précision des inférences. Cette combinaison permet de tirer parti à la fois de la structure formelle des ontologies et des capacités d'apprentissage automatique des plongements de graphes.

Nous illustrons ces méthodes de fouille de graphes avec deux exemples tirés de la littérature.

MONNIN et al. [134] utilisent les Graph Convolutional Networks (GCN), un algorithme de plongement de graphes, pour réaliser l'alignement ontologique dans PGxLOD. Les GCN

appliquent des opérations de convolution sur les entités et leurs voisins pour produire des représentations vectorielles. Les résultats montrent que les GCN peuvent efficacement capturer les similitudes structurelles et sémantiques entre les entités de différentes ontologies, facilitant ainsi l'alignement ontologique. Dans le cas de PGxLOD, cette technique permet d'enrichir le graphe et d'augmenter la confiance en certaines unités pharmacogénomiques (association gène-médicament-phénotype) : si une unité est retrouvée dans différentes ontologies grâce aux GCN, cela renforce la validité ou la pertinence de cette association.

DRANCE et al. [53] utilisent des plongements de graphes pré-entraînés, en particulier l'algorithme ConvE (*Convolutional Embedding*), pour améliorer le raisonnement multi-saut (*multi-hop reasoning*). ConvE applique des opérations de convolution sur les graphes pour apprendre des représentations vectorielles des entités et des relations. Le raisonnement multi-hop consiste à faire des inférences à travers plusieurs étapes de relations dans le graphe. Ils évaluent l'efficacité de leur approche sur des tâches de prédiction de liens en utilisant des ensembles de données de référence. L'intégration des plongements de graphes avec le raisonnement multi-saut permet d'améliorer la précision et l'efficacité des inférences dans les graphes de connaissances, et pourrait être appliquée à des graphes de connaissances bio-médicaux.

2.5 État de l'art des modèles de données spécifiques aux chimiothérapies

Les schémas thérapeutiques et les réponses aux chimiothérapies sont des objets complexes et de ce point de vue la peuvent tirer parti de représentation de connaissances. Plusieurs représentations ont été proposées dans la littérature pour représenter connaissances et données associées. Nous en présentons les principaux dans cette section.

Le modèle de données Observational Medical Outcomes Partnership Common Data Model (**OMOP CDM**) [151] est un schéma relationnel standardisé conçu pour harmoniser les données cliniques à des fins de recherche en santé. Il se distingue du modèle en étoile d'i2b2 [94] (cf. 1.4.4) par un plus grand nombre de tables pour décrire les données observées. Chacune de ces tables est dédiée à un aspect spécifique des données cliniques. Par exemple, la table "PERSON" contient des informations démographiques de chaque patient, tandis que "DRUG_EXPOSURE" capture les expositions aux médicaments, incluant la dose et la durée d'exposition. Ce modèle permet d'intégrer à la fois des données cliniques et médico-administratives. Une table clé d'OMOP est "Standardized vocabularies", qui lie les concepts médicaux à des terminologies de référence, et assure ainsi une interopérabilité et une standardisation des données.

L'extension OMOP CDM pour le cancer [16] enrichit ce modèle avec des tables supplémentaires pour capturer des données spécifiques au cancer, et permet d'inclure des détails sur les traitements de chimiothérapie, les réponses aux traitements et les caractéristiques des tumeurs. Les tables "EPISODE" et "EPISODE_EVENT" permettent notamment de capturer le cours du traitement et les événements de réponses. La table "EPISODE" est hiérarchiquement imbriquée : elle peut inclure d'autres enregistrements de la table "EPISODE". Par exemple, un épisode de cancer du sein peut contenir plusieurs sous-épisodes correspondant

aux différents stades de la maladie ou aux différentes lignes de traitement qui elles-mêmes peuvent contenir des sous-épisodes correspondant aux cycles. Ceci permet une représentation fidèle des parcours de soin et de leur temporalité. Chaque épisode peut être lié à des événements cliniques spécifiques, comme la survenue d'une toxicité, via la table "EPISODE_EVENT".

Parmi les terminologies de référence de la table "Standardized vocabularies" d'OMOP, WARNER et al. [212] montrent qu'[HemOnc](#) peut être utilisée pour standardiser les schémas thérapeutiques de chimiothérapie décrivant les cycles des lignes suivies par les patients. [HemOnc](#) [213], est une terminologie spécialement conçue pour représenter les schémas thérapeutiques de chimiothérapie. Elle en fournit une description détaillée, en décrivant la durée du cycle et la combinaison d'administrations de médicaments anticancéreux, ainsi que les toxicités fréquemment observées.

Le modèle OMOP CDM offre une grande richesse dans sa représentation des chimiothérapies et de leurs réponses, en permettant à la fois de lier les chimiothérapies suivies aux schémas thérapeutiques grâce à la standardisation avec HemOnc, de représenter la temporalité du traitement avec l'imbrication de la table EPISODE, et d'y associer des réponses avec la table EPISODE_EVENT. De plus, le modèle théorique d'OMOP a été instancié dans de nombreux hôpitaux, avec plusieurs applications de recherche sur le cancer, ce qui témoigne de sa robustesse et de son utilité pratique [3]. Cependant, pour illustrer la pertinence des ontologies dans la représentation des chimiothérapies et de leurs réponses, il est essentiel d'examiner également les modèles ontologiques présents dans la littérature.

L'ontologie Cancer Care Treatment Outcome Ontology ([CCTOO](#)) [115] est conçue pour représenter les résultats des traitements chez les patients atteints de tumeurs solides. CCTOO a été instanciée à partir des données de [ClinicalTrials.gov](#), une base de données d'essais cliniques qui fournit des informations détaillées et structurées sur les schémas thérapeutiques à l'essai et les réponses aux traitements (cf. sous-section 1.3.1). Cette source de données, bien que riche et détaillée, introduit un biais car les données issues des essais cliniques sont souvent plus structurées et plus complètes que les données de vie réelle (cf. sous-section 1.5.1). L'ontologie CCTOO réutilise des terminologies de référence telles que SNOMED, NCIt et UMLS pour capturer une large gamme de concepts cliniques et permet de décrire de manière détaillée l'histoire des traitements des patients et leurs réponses. Par exemple, elle inclut des concepts détaillés pour la réponse tumorale, la survie sans progression, et les événements indésirables graves, offrant ainsi une représentation riche et exhaustive des réponses aux traitements de chimiothérapie. Cependant, elle manque de détails sur la représentation de la temporalité des traitements.

Par contraste, CHEN et al. [35] utilisent la Time Event Ontology ([TEO](#)) [113], une ontologie spécifiquement conçue pour représenter les événements médicaux, leur temporalité, et le cours des traitements de cancer. Instanciée à partir des données de vie réelle du Cancer Data Standards Repository ([caDSR](#)), cette ontologie permet de modéliser précisément la séquence et la durée des événements médicaux. Elle inclut des concepts pour représenter les intervalles de temps, les points temporels et les relations temporelles entre les événements, ce qui en fait un outil puissant pour l'analyse temporelle des données cliniques. Cependant,

bien que TEO soit conçue pour capturer la temporalité, elle ne fournit pas une description détaillée des protocoles de chimiothérapie sous forme de séquences de cycles, ce qui limite son application pour représenter les lignes de traitement de chimiothérapie de manière complète. De plus, nous verrons dans la sous-section 2.3.4.1 qu'il existe une ontologie de référence pour représenter le temps, la Time Ontology.

Chaque modèle de données présenté ici offre des avantages uniques pour la représentation des chimiothérapies et de leurs réponses. OMOP CDM, avec son extension pour le cancer et son intégration de HemOnc, fournit une structure relationnelle robuste pour capturer les traitements de chimiothérapie et leurs réponses, avec une standardisation et une interopérabilité élevées. Les ontologies CCTOO et TEO offrent des représentations détaillées respectivement des réponses aux traitements et de la temporalité des événements médicaux, bien qu'elles soient respectivement limitées par l'absence de données de vie réelle et la représentation complète des lignes suivies et des schémas thérapeutiques. Une approche intégrée combinant les éléments considérés indépendamment par ces modèles pourrait offrir une solution plus complète pour la recherche en oncologie.

2.5.1 Note sur les représentations temporelles de la TEO et de la TO

Bien que la Time Ontology (TO) 2.3.4.1 permette une description précise de la temporalité, son adoption dans les modèles biomédicaux reste limitée. Dans la section précédente, on a montré comment CHEN et al. [35] représentaient les événements de chimiothérapies avec la Time Event Ontology (TEO) [113] (anciennement Clinical Narrative Time Relational Ontology - CNTRON [198, 199]). La TEO est une ontologie spécifiquement conçue pour représenter la temporalité des événements médicaux à partir d'extractions de dossiers patients. Cette ontologie utilise donc sa propre représentation du temps, et ne réutilise pas la TO. Chaque extraction d'intérêt associée à une temporalité dans les dossiers patients instancie une sous-classe de la classe "TEO :Event". BATSAKIS et al. [15] exposent la différence de représentation de leur modèle 4D-fluent avec la TO, et le modèle de représentation de la TEO qui selon eux est un modèle de réification. Avec le modèle de réification, une entité est générée pour chaque association à temporaliser tandis qu'avec le modèle 4D-fluent, une entité temporelle est générée pour chacune des entités impliquées dans une association à temporaliser. La figure 2.1, tirée de l'article de BATSAKIS et al. [15], illustre les deux modèles avec un exemple non médical.

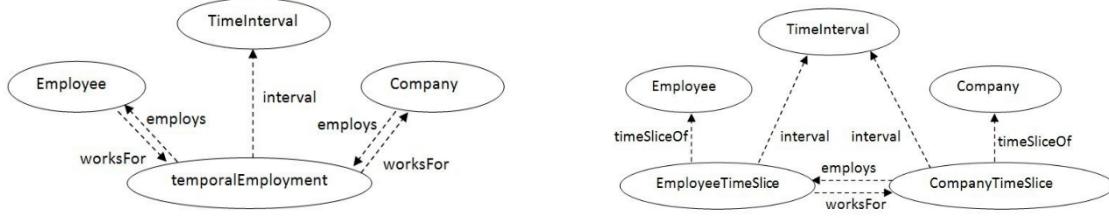


FIGURE 2.1 – Modélisation de la temporalité d'un contrat de travail entre une entreprise et un salarié avec les modèles de réification et 4D-fluent. À gauche le modèle de réification utilisé dans la Time Event Ontology. À droite le modèle 4D-fluent modélisable avec la TO. La figure est tirée de l'article "Temporal representation and reasoning in OWL 2" de BATSAKIS et al. [15].

2.6 Bilan : pourquoi les ontologies ?

Les ontologies semblent avoir un potentiel intéressant pour représenter les connaissances liées aux chimiothérapies et leurs réponses, grâce à leur capacité à structurer, intégrer et interconnecter des données complexes provenant de sources hétérogènes. Elles permettent de surmonter les limites des entrepôts de données traditionnels en offrant une représentation standard et extensible des connaissances biomédicales.

Les ontologies fournissent un cadre interopérable qui facilite la communication et l'échange de données entre différents systèmes et institutions. En utilisant les technologies du Web sémantique les ontologies facilitent la réutilisation de sources déjà disponibles et s'appuient sur des connaissances de domaines déjà formalisées, en plus d'observer des données.

Les ontologies permettent également de réaliser des inférences et des raisonnements complexes, bien que les limites computationnelles des raisonneurs imposent parfois des défis. Pour pallier ces défis, l'utilisation de techniques de fouille de graphes, telles que les plongements de graphes, s'avère particulièrement prometteuse. Ces méthodes permettent d'identifier des motifs récurrents, de découvrir des relations cachées et d'améliorer la qualité des connaissances intégrées.

Dans le prochain chapitre, nous nous pencherons sur les techniques de détection et de prédiction d'événements à partir de données médicales. Nous explorerons comment extraire des informations pertinentes de diverses sources de données, que ce soit les données brutes des DPI à nettoyer avant d'instancier les ontologies, ou les données structurées des bases de connaissances. Nous verrons comment ces informations peuvent être utilisées pour enrichir les bases de connaissances et soutenir des analyses plus approfondies, telles que la prédiction de la réponse aux traitements.

CHAPITRE 3

EXTRACTION DE CONNAISSANCES ET PRÉDCTION D'ÉVÉNEMENTS

3.1	Extraction de connaissances	60
3.1.1	Extraction de connaissances à partir du texte libre	60
3.1.2	Des mesures pour extraire des connaissances	63
3.2	Prédiction d'événements	67
3.2.1	Analyse de survie classique	67
3.2.2	Régression logistique pour l'analyse de survie	81
3.2.3	Comparaison des modèles de survie étudiés	84
3.2.4	Prédictions d'événements avec des modèles de survie	86
3.3	Bilan : quels outils pour extraire et prédire les réponses aux chimiothérapies ?	91



Nous avons vu dans le chapitre précédent, que les ontologies permettaient la représentation d'un domaine médical spécifique, l'intégration des connaissances sur ce domaine et leur exploration. Ces connaissances sont dispersées dans diverses sources dans les entrepôts de données. Ces sources sont structurées (résultats biologiques, codage d'actes ou de diagnostiques), semi-structurées (réponses à des questionnaires, tableaux des comptes rendus, critères RECIST et TNM comptes rendus), ou non structurées (texte libre dans les comptes rendus). Dans de nombreux cas, les données brutes ne sont pas directement exploitables. Ces données nécessitent d'être traitées pour en extraire des connaissances.

Dans ce chapitre, nous présentons quelques techniques d'extraction des connaissances à partir des données bio-médicales. Nous nous intéressons ensuite à la prédiction d'événements à partir des connaissances extraites. Il est divisé en deux sections.

La première section est dédiée à l'extraction des connaissances. (section 3.1) Premièrement nous présentons quelques techniques de traitement automatique des langues (TAL) permettant d'extraire des connaissances à partir du texte libre. Nous nous focalisons sur deux tâches (sous-section 3.1.1) :

- la reconnaissance d'entités nommées (REN) et de leur contexte.
- l'extraction de relation (ER) entre entités d'intérêt.

Dans un deuxième temps, nous nous concentrerons sur la présentation de mesures pour extraire des connaissances sur un grand volume de données. (section 3.1.2)

Dans la deuxième section, nous nous intéressons à la prédition d'événements (section 3.2). Cette section est composée de quatre sous-sections.

- Tout d'abord, nous présentons l'analyse de survie classique. Nous illustrons son application pour étudier des facteurs de risques de survenue d'événements avec deux études de cas. (section 3.2.1)
- Puis, nous nous intéressons à l'utilisation de la régression logistique pour analyser la survie avec des modèles de temps continu ou discret. (section 3.2.2)
- Ensuite nous comparons les différents modèles de survie exposés, et leurs différentes interprétations. (section 3.2.3)
- Enfin nous nous intéressons à l'utilisation des modèles de survie pour prédire la survenue d'événements. Nous expliquons comment développer des modèles prédictifs à travers la construction de nomogrammes. (section 3.2.4)

La section sur l'analyse de survie classique (section 3.2.1) s'appuie sur le [tutoriel en ligne](#) de Lisa Sullivan, Professeure de Biostatistique à la Boston University School of Public Health et sur la thèse (en cours d'écriture) [162] de [Juliette Murris](#) [137].

3.1 Extraction de connaissances

Définition - Extraction de connaissances

L'extraction de connaissances^a est le processus de création de connaissances à partir d'informations structurées, semi-structurées ou non structurées. Le résultat doit être dans un format lisible par les ordinateurs.

a. FAYYAD, PIATECKY-SHAPIRO et SMYTH [63]

3.1.1 Extraction de connaissances à partir du texte libre

L'extraction de connaissances à partir du texte implique l'utilisation d'outils de Traitement Automatique des Langues (TAL ou NLP pour *Natural Language Processing*). Le TAL couvre de nombreuses tâches et connaît des développements fulgurants. Cette section n'a pas pour objectif d'en faire un état de l'art exhaustif. Nous nous concentrerons sur les tâches que nous avons utilisées dans nos contributions, qui reposent sur des méthodes classiques assez éloignées des méthodes de pointe actuelles. Parmi les tâches de TAL, nous nous intéresserons particulièrement à la reconnaissance d'entités nommées (REN ou NER pour *Named-Entity Recognition*) à base de règles et de dictionnaires, à l'extraction de leur contexte et à l'extraction de relations (ER).

3.1.1.1 Reconnaissance d'Entités Nommées

La Reconnaissance d'Entités Nommées (REN) est une tâche de TAL visant à identifier et classifier des entités textuelles telles que des noms de personnes, d'organisations, de lieux, de dates, etc [80]. Les approches REN peuvent être classées en trois catégories : les approches à base de règles, les approches basées sur des dictionnaires, et les approches supervisées.

Approches basées sur des dictionnaires

Les approches basées sur des dictionnaires utilisent des dictionnaires pour identifier les entités dans le texte. Ces méthodes s'appuient sur :

- un dictionnaire,
- un algorithme de parcours de textes d'un corpus et de parcours du dictionnaire,
- un algorithme pour calculer les distances entre mots du dictionnaire et du texte (*cf.* les distances d'édition définies dans la section suivante).

Les dictionnaires contiennent un ensemble de termes pour identifier les identités d'intérêt. Ces approches sont assez courantes dans le domaine biomédical qui, comme on l'a montré dans le chapitre précédent 2.5, dispose de nombreuses ontologies/terminologies et classifications exploitables pour construire un dictionnaire pour la REN.

Par exemple, QuickUMLS [192] est un outil qui permet la REN avec des dictionnaires liés aux concepts de l'UMLS, facilitant ainsi la normalisation des extractions et leur intégration au sein d'un arbre de concepts. Pour réaliser cette correspondance rapide et approximative des chaînes de caractères, QuickUMLS utilise Simstring [152]. Simstring peut employer différentes mesures de similarité, pour comparer les chaînes de caractères (*cf.* section 3.1.2). QuickUMLS est particulièrement utile pour l'extraction d'entités médicales en raison de sa tolérance aux variations orthographiques d'une part, et de ses liens vers l'UMLS.

L'outil IAM system, [43], utilise également des approches basées sur des dictionnaires pour la reconnaissance d'entités nommées, en employant la distance de Levenshtein (*cf.* section 3.1.2) pour mesurer la similarité entre les termes. COSSIN et al. [42] utilisent IAM system pour extraire les médicaments de comptes rendus médicaux.

Approches à base de règles

Les approches à base de règles reposent sur des ensembles de règles prédéfinies pour identifier les entités nommées. Ces règles peuvent inclure des motifs lexicaux ou des expressions régulières. EFTIMOV, KOROUŠIĆ SELJAK et KOROŠEC [59] utilisent une approche à base de règles pour extraire des entités nutritionnelles à partir de textes libres. Cependant, ces approches sont souvent critiquées pour leur manque de généralisation, car elles sont souvent spécifiques au domaine étudié. En effet, l'ensemble des règles est développé et affiné pour une application particulière et sont difficilement réutilisables pour une autre application, ou un autre corpus.

Approches supervisées Les approches supervisées utilisent des algorithmes d'apprentissage automatique, tels que les réseaux de neurones ou les machines à vecteurs de support (SVM pour *Support Vector Machine*), pour prédire les entités dans un texte. Par exemple, les approches neuronales pour la REN exploitent des réseaux de neurones pour apprendre des représentations vectorielles des mots, ce qui permet d'identifier et de classifier automatiquement les entités nommées à partir de données annotées [39]. LERNER, PARIS et TANNIER [112] décrivent l'utilisation de réseaux de neurones récurrents (*RNN* pour *Recurrent Neural*

Network) pour l'extraction d'entités nommées dans le domaine médical, démontrant que les approches neuronales peuvent améliorer la précision de l'extraction par rapport aux méthodes traditionnelles.

3.1.1.2 Extraction du contexte de l'entité

Pour interpréter correctement une extraction d'entité, il faut aussi extraire son contexte. Sans cela, la détection d'une entité peut entraîner la détection de faux positifs. La détection du contexte inclut par exemple la détection de la négation, des hypothèses, des antécédents familiaux et de la temporalité. GARCELON et al. [71] utilisent une approche à base de règles pour détecter le contexte des entités extraites dans les comptes rendus médicaux. Le système de règles permet d'identifier si une entité mentionnée dans le texte est présente, absente, ou hypothétique, améliorant ainsi la précision et la pertinence des informations extraites. En revanche, il ne permet pas de détecter la temporalité de l'extraction. La détection de la temporalité est assez compliquée car elle suppose souvent d'extraire des entités de temporalité, qui peuvent être des dates ou des expressions, et de les lier à une portion de texte à laquelle elle s'applique. BANNOUR et al. [14] proposent un modèle neuronal pour identifier les entités temporelles et les lier à des entités d'intérêt.

3.1.1.3 Extraction de relations

L'extraction de relations vise à identifier et à classifier les relations entre les entités nommées dans le texte [63].

Établir une relation entre deux entités avec la co-occurrence

La co-occurrence est une méthode pour extraire des relations en comptant le nombre de fois où deux entités apparaissent ensemble dans une même portion de texte. AVILLACH et al. [10] utilisent cette approche pour détecter des relations entre les événements indésirables et les médicaments dans les dossiers de patients.

Établir une relation entre deux entités avec un parseur de dépendance

Les parseurs de dépendance analysent la structure syntaxique des phrases pour identifier les relations grammaticales entre les mots. HASSAN et al. [84] utilisent un parseur de dépendance pour établir des liens entre des entités de symptômes et de maladies extraites par des techniques de REN basées sur des dictionnaires. Les parseurs de dépendance permettent d'établir si deux entités identifiées dans une phrase sont syntaxiquement liées, ce qui améliore la précision de l'extraction des relations. Les figures 3.1 et 3.2 illustrent ce processus.

Ex 4.1. “A 15-month-old girl with < disease > propionic acidemia </ disease > presented < symptom > muscular hypotonia </ symptom >”

Ex 4.2. “A 25-year-old woman with < disease > cystic fibrosis </ disease > developed < symptom > hemoptysis </ symptom >”

FIGURE 3.1 – Reconnaissances des entités d'intérêts maladie et symptôme dans la même phrase.

FIGURE 3.3 – Utilisation des parseurs de dépendance pour l'extraction de relations entre les entités d'intérêt maladie et symptôme. Figures tirés de l'article de HASSAN et al. [84].

Nous avons présenté des outils de TAL permettant d'extraire des entité avec leur contexte depuis le texte libre et d'établir des relations entre ces entités. Nous utilisons ces méthodes pour extraire les toxicités du texte libre (*cf. chapitre 4*). Dans la section suivante, nous nous intéressons aux mesures permettant de détecter du signal sur un grand volume de données.

3.1.2 Des mesures pour extraire des connaissances

Définition - Mesure

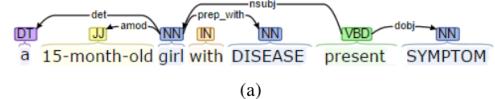
Une mesure^a est une fonction qui associe une grandeur numérique à certains sous-ensembles d'un ensemble donné.

a. WIKIPÉDIA [218]

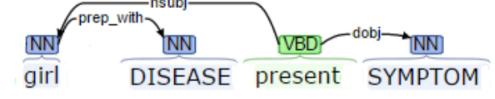
Dans cette section, nous explorons l'utilisation des mesures dans des grands ensembles de données pour extraire des connaissances. Nous nous concentrerons dans un premier temps sur les mesures de similarité pour comparer des chaînes de caractères, une technique fondamentale du traitement automatique des langues (TAL), et nous montrerons leurs applications dans des domaines variés. Ensuite, nous nous intéresserons à des mesures cliniques utilisées à grande échelle dans la recherche pour vérifier des hypothèses.

Cette section sur les mesures est motivée par le développement d'une mesure, Proto-Drift, parmi les contributions de cette thèse.

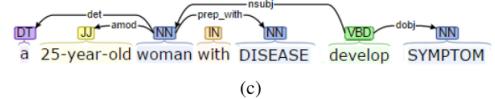
3.1.2.1 Comparer des chaînes de caractères avec des mesures de similarité



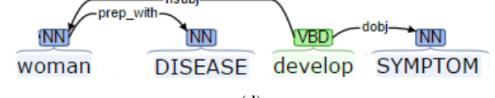
(a)



(b)



(c)



(d)

FIGURE 3.2 – Liaison des entités d'intérêt avec le parseur de dépendance.

FIGURE 3.3 – Utilisation des parseurs de dépendance pour l'extraction de relations entre les entités d'intérêt maladie et symptôme. Figures tirés de l'article de HASSAN et al. [84].

Beaucoup de mesures de similarité sont utilisées pour extraire des connaissances. Avant de montrer des exemples d'utilisation de ces mesures, nous définissons les notions de distance, similarité, dissimilarité et de mesure de similarité.

Définition - Distance

On appelle distance^a sur un ensemble E toute application d définie sur le produit $E^2 = E \times E$ et à valeurs dans l'ensemble \mathbb{R}^+ des réels positifs ou nuls,

$$d : E \times E \rightarrow \mathbb{R}^+$$

vérifiant les propriétés suivantes :

- **Symétrie** : $\forall (a, b) \in E^2, d(a, b) = d(b, a)$
- **Séparation** : $\forall (a, b) \in E^2, d(a, b) = 0 \iff a = b$
- **Inégalité triangulaire** : $\forall (a, b, c) \in E^3, d(a, c) \leq d(a, b) + d(b, c)$

Un ensemble E muni d'une distance d s'appelle un espace métrique.

a. WIKIPÉDIA [217]

Par exemple, la distance euclidienne entre deux points $a = (a_1, a_2)$ et $b = (b_1, b_2)$ dans un espace à deux dimensions \mathbb{R}^2 est définie par :

$$d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

La distance euclidienne est une application à valeurs dans \mathbb{R}^+ qui vérifie les trois propriétés de symétrie, séparation et d'inégalité triangulaire.

Définition - Similarité et dissimilarité

La similarité entre deux entités a et b notée $sim(a, b)$ est le degré de ressemblance entre deux entités, indiquant jusqu'à quel point elles sont semblables. La similarité n'a pas de définition mathématique formelle. C'est une fonction qui prend deux entités en arguments, et qui augmente avec le degré de ressemblance entre ces deux entités. Le complémentaire de la similarité, la dissimilarité, notée $1 - sim(a, b)$, quantifie le degré de différences entre deux entités.

Par exemple, la similarité cosinus entre deux vecteurs a et b dans un espace à deux dimensions \mathbb{R}^2 est définie par :

$$sim(a, b) = \frac{a \cdot b}{\|a\| \|b\|}$$

où $a \cdot b$ est le produit scalaire des vecteurs a et b , et $\|a\|$ et $\|b\|$ sont les normes euclidiennes de a et b , respectivement.

La similarité cosinus augmente à mesure que les vecteurs a et b deviennent plus alignés dans la même direction.

Les dissimilarités ne sont pas nécessairement des distances. Par exemple, la dissimilarité cosinus (complémentaire de la similarité cosinus) ne satisfait pas l'inégalité triangulaire, et n'est donc pas une distance [182].

Définition - Mesure de similarité

Les mesures de similarité mesurent des similarités ou des dissimilarités (qui peuvent être des distances).

Dans les approches dictionnaires de la REN (*cf.* section 3.1.1.1.1), nous avons vu que des mesures de similarité sont utilisées pour comparer deux chaînes de caractères. Ces mesures de similarité sont souvent des distances, et dans ce cas, on les appelle des distances d'édition.

Définition - Distance d'édition

Soit Σ un alphabet, c'est-à-dire un ensemble fini de caractères. Soit Σ^* l'ensemble de toutes les chaînes de caractères formées par l'alphabet Σ . Chaque élément de Σ^* est une chaîne de caractères de longueur finie sur Σ . Par exemple, si $\Sigma = \{a, b\}$, alors $\Sigma^* = \{a, b, aa, ab, ba, bb, aaa, \dots\}$. Une distance d'édition sur Σ^* est une application d sur $\Sigma^* \times \Sigma^*$ et à valeurs dans l'ensemble \mathbb{R}^+ des réels positifs ou nuls, qui vérifie les trois propriétés de symétrie, de séparation et d'inégalité triangulaire [143].

Le principe des distances d'édition est de compter le nombre minimal d'opérations basiques nécessaires pour passer d'une chaîne à l'autre. Quatre catégories d'édition existent :

- la substitution d'un caractère : "livre" en "libre" ;
- l'effacement d'un caractère : "porte" en "port" ;
- l'insertion d'un caractère : "main" en "matin" ;
- la transposition : "chien" en "chine".

Les différentes distances d'édition existantes comptent différents ensembles d'opérations. Par exemple, la distance de Levenshtein compte le nombre de substitutions, d'effacements et d'insertions, tandis que la distance de Hamming, définie pour des chaînes de caractères de même longueur, compte seulement le nombre de substitutions.

L'approximation de chaîne de caractères dans les dictionnaires (*cf.* section 3.1.1.1.1) peut être effectuée via cet algorithme [143] :

- Soit $T \in \Sigma^*$ un texte de longueur $n = |T|$.
- Soit $P \in \Sigma^*$ un motif de longueur $m = |P|$.
- Soit $k \in \mathbb{R}^+$ le nombre maximal d'erreurs autorisées.
- Soit $d : \Sigma^* \times \Sigma^* \rightarrow \mathbb{R}^+$ une fonction de distance d'édition.

Étant donné T , P , k et $d(\cdot)$, retourner l'ensemble de toutes les positions j dans le texte telles qu'il existe un i pour lequel $d(P, T_{i..j}) \leq k$.

En bio-informatique on utilise des mesures de similarité pour comparer des séquences biologiques telles que des séquences génomiques ou des protéines. L'ordre des nucléotides dans les gènes et des acides aminés dans les protéines déterminent une partie de leurs fonctions biologiques. Ainsi deux séquences sont dites similaires si l'ordre des caractères (nucléotides ou acides aminés) qui les composent est conservé. Ces séquences peuvent être représentées avec des chaînes de caractères. La comparaison de séquences biologiques revient

donc alors à une comparaison de chaîne de caractères. En 1970, NEEDLEMAN et WUNSCH [144] proposent un algorithme d’alignement de séquence dont l’objectif est de maximiser une similarité en appariant le maximum de caractères ordonnés. Pour favoriser l’appariement de caractères ordonnés, l’algorithme d’alignement de séquence autorise une opération supplémentaire par rapport aux opérations des distances d’édition classiques : les insertions d’espaces vides (*gaps*). La similarité produite par l’algorithme est un entier relatif et plus il est élevé, plus les séquences se ressemblent. L’algorithme de NEEDLEMAN et WUNSCH [144] aligne globalement deux séquences biologiques. Quelques années plus tard, SMITH, WATERMAN et al. [188] proposent un algorithme d’alignement semblable mais qui favorise les alignements locaux. SELLERS [185] montre que la maximisation de la similarité des algorithmes d’alignement biologiques peut être ramené à une minimisation d’une distance d’édition, qui autorise les insertions d’espaces vides (*gaps*). Ces algorithmes de comparaison de chaînes de caractères sont des algorithmes de programmation dynamique, que nous ne détaillerons pas.

Les algorithmes d’alignements biologiques permettent, entre autre, d’étudier les fonctions des gènes, prédire la structure des protéines pour le développement de médicaments, et en phylogénie ils permettent d’étudier l’évolution des espèces [9]. Au-delà des alignements de séquences biologiques, ce qui est intéressant avec ces algorithmes, c’est qu’ils mettent l’accent sur la séquence des caractères d’une chaîne de caractères. En informatique médicale, plusieurs études s’en inspirent pour aligner les trajectoires de maladies de patients [75]. Les chaînes de caractères sont des séquences d’événements médicaux codés, par exemple avec les codes diagnostiques de la classification internationale des maladies (CIM, *ICD*, cf. table 2.1), et on étudie les similarités de trajectoires. Les protocoles théoriques et les lignes suivies de chimiothérapies peuvent aussi être représentés comme des séquences d’événements. Ainsi, nous nous sommes inspirés des alignements de séquences pour développer ProtoDrift (cf. chapitre 5).

On a montré des exemples d’utilisation de mesures de similarité pour extraire des connaissances, et montrer leur application dans divers domaines. Dans la section suivante, nous nous intéressons aux mesures utilisées dans la recherche clinique.

3.1.2.2 Mesures cliniques utilisées à grande échelle dans la recherche clinique

Dans de nombreuses études cliniques, on évalue la corrélation entre deux mesures, ou entre une mesure clinique et la survenue d’un événement pour tester des hypothèses.

Par exemple SCHUURHUIZEN et al. [183] mesurent à la fois un score de toxicité cumulée et la qualité de vie des patients avec le questionnaire de référence EORTC-QLQ-C30 [166] chez des patients atteints de cancers colorectaux. Ils évaluent différents paliers de grades de toxicité dans le score de toxicité cumulé et leurs associations avec la qualité de vie des patients. Ils montrent que la prise en compte de tous les grades dans le score de toxicité cumulée est significativement corrélée à la qualité de vie physique des patients. Ils concluent que les futurs essais cliniques devraient prendre en compte tous les grades de toxicité pour évaluer la qualité de vie des patients.

Un autre exemple de mesure utilisée dans les recherches cliniques que nous avons documenté dans le chapitre 1 est la RDI. On l'utilise pour mesurer l'adhésion à la chimiothérapie. On fixe un seuil et on évalue la survie des patients stratifiés de part et d'autre de ce seuil. On peut comparer la RDI à la mesure normalisée de la coagulation sanguine (INR pour *International Normalised Ratio*). L'INR est utilisé pour mesurer le temps de coagulation sanguine et ajuster les doses d'anticoagulants. L'objectif est de maintenir l'INR dans une plage thérapeutique cible, généralement entre 2 et 3 pour de nombreux traitements, afin de prévenir les complications hémorragiques ou thrombotiques.

Notons que l'INR et la RDI ne sont pas des distances (*cf.* définition de la distance 3.1.2.1). En effet, elles ne vérifient pas la propriété de symétrie. Ceci est directement lié à leur objectif qui est de mesurer un écart par rapport à une référence, pour mesurer une adhésion. Ce sont des ratios normalisés.

Comme pour la RDI, l'INR est comparé à une référence pour évaluer son efficacité. De même que pour les études utilisant la RDI, les patients sont stratifiés en fonction de leurs valeurs d'INR, et leur survie est analysée pour déterminer l'impact de l'adhésion au traitement anticoagulant sur les résultats cliniques [148].

Ceci nous amène à la section suivante où nous présentons la mise en œuvre d'une analyse de survie.

3.2 Prédiction d'événements

3.2.1 Analyse de survie classique

L'analyse de survie est une méthode statistique utilisée pour étudier le temps écoulé avant la survenue d'un événement d'intérêt. En santé, cette analyse est massivement utilisée pour examiner les facteurs de risques contribuant à la survenue d'événements cliniques importants [38], telles que la rémission d'une maladie ou le décès.

En oncologie, les événements d'intérêt incluent souvent la rechute du cancer, la progression de la tumeur et le décès. Les principales mesures de survie utilisées sont les suivantes [47] :

- **La survie sans récidive (SSR ou Disease-Free Survival, DFS)** : Lorsqu'il s'agit de la rémission de la maladie, telle que l'absence de rechute du cancer après un traitement initial. La SSR mesure le temps écoulé entre le traitement curatif et la première réapparition de la maladie ou le décès dû à toute cause.
- **La survie sans progression (SSP ou Progression-Free Survival, PFS)** : Lorsque l'événement étudié concerne la progression de la maladie, comme l'augmentation de la taille de la tumeur ou l'apparition de nouvelles lésions. La SSP mesure le temps écoulé entre le début du traitement et la progression de la maladie ou le décès dû à toute cause. La progression est généralement évaluée à l'aide des critères RECIST, qui permettent de mesurer objectivement les changements dans la taille des tumeurs et de détecter de nouvelles lésions (*cf.* section 1.2.3).
- **La survie globale (SG ou Overall Survival, OS)** : Lorsque l'événement d'intérêt est

le décès du patient, quelle qu'en soit la cause. La SG mesure le temps écoulé entre le début du suivi (par exemple, le diagnostic ou le début du traitement) et le décès.

Ces mesures de survie permettent de comprendre et évaluer l'efficacité des traitements, comme nous allons le voir dans les sections suivantes, mais aussi de prédire les résultats cliniques pour les patients atteints de cancer (*cf. section 3.2.4*).

Au cours du chapitre 1, nous avons vu que des analyses de survie étaient utilisées pour tester les options de schémas thérapeutiques (*cf. section 1.3.1*), et pour évaluer l'impact de l'adhésion thérapeutique avec la dose-intensité relative (RDI) (*cf. section 1.3.2*).

3.2.1.1 Objectifs de l'analyse de survie classique

Les analyses de survie permettent d'évaluer l'impact de variables d'intérêt, souvent appelées facteurs de risque, sur la survenue d'un événement. En général, ces analyses fournissent des réponses avec trois approches :

1. La visualisation graphique des estimations de courbes de survie pour des groupes définis selon la variable d'intérêt à l'aide de l'estimateur de fonction de survie de Kaplan-Meier (*cf. section 3.2.1.7*)
2. La comparaison statistique des estimations de fonctions de survie entre ces groupes à l'aide du test du log-rank (*cf. section 3.2.1.8*)
3. L'analyse de l'impact de la variable d'intérêt et d'autres variables explicatives sur le risque de survenue de l'événement en ajustant des modèles de Cox et en évaluant la significativité des coefficients de ces variables (*cf. section 3.2.1.10*)

Nous illustrons la mise en oeuvre de l'analyse de survie avec ces trois approches avec deux études.

Dans la première, MENG et al. [131] étudient l'impact sur la survie sans progression (SSP) et la survie globale (SG) de l'ajout d'une thérapie ciblée (bévacizumab) en première ligne de traitement chez des patients atteints de cancer colorectal avancé et muté. Deux groupes de patients sont distingués selon qu'ils reçoivent ou non la thérapie ciblée en première ligne. Les analyses de survie sont menées avec la progression de la tumeur (SSP) et le décès (SG) comme événements d'intérêts, et l'option avec ou sans thérapie ciblée comme une des variables explicatives, encodée de façon binaire. Les deux groupes sont comparés par leurs estimations de Kaplan-Meier, les résultats du test du log-rank, et l'analyse de la contribution des variables explicatives dans des modèles de Cox.

Dans la seconde étude, BREADNER et al. [24] examinent l'impact des RDI de trois anticancéreux, l'oxaliplatin, le 5-FU et la capécitabine, sur la SG chez des patients atteints de cancer du colon. Le patients doivent avoir reçu de l'oxaliplatin en combinaison soit avec le 5-FU soit avec la capécitabine.

Pour chacune des trois molécules :

- des groupes de patients sont stratifiés en fonction de la RDI reçue ($\leq 80\%$ vs. $> 80\%$);
- des estimations de Kaplan-Meier des courbes de survie sont réalisées sur ces groupes

- les fonctions de survie sont comparées avec des tests de log-rank

Un modèle de Cox est évalué avec les RDI des trois molécules parmi les variables explicatives, encodées de façon binaire ($\leq 80\%$ vs. $> 80\%$). Les contributions des ces variables sont évaluées.

Dans les sections suivantes, nous présentons les bases méthodologiques de ces trois approches, en illustrant leur application à ces deux études.

3.2.1.2 Définitions des dates en analyse de survie

Plusieurs dates clés sont utilisées pour mesurer le temps écoulé jusqu'à l'événement d'intérêt :

- **Date d'origine (DO)** : Le point de départ pour le suivi de chaque patient, qui peut être la date de randomisation dans un essai clinique, la date d'inclusion dans une étude de cohorte, ou la date de diagnostic. Cette date varie pour chaque patient.
- **Date de dernières nouvelles (DDN)** : La dernière date à laquelle des informations fiables sur le patient sont disponibles. Cette date varie également pour chaque patient.
- **Date de point (DP)** : Cette date est déterminée a priori, souvent dans le protocole de l'étude, et sert de repère pour l'analyse des données. Elle correspond à la date de l'analyse intermédiaire ou finale de l'étude. Après cette date, les informations supplémentaires éventuellement collectées ne sont généralement pas prises en compte dans l'analyse principale. Cette date est commune à tous les patients.

Les deux études que nous avons prises pour exemples sont des études rétrospectives. MENG et al. [131] ont inclus les données de patients débutant leur première ligne de chimiothérapie entre mars 2015 et août 2021. La DO correspond à la date de début de première ligne pour chaque patient, comprise entre le 01/03/2015 et le 31/08/2021. La DDN varie pour chaque patient et est également comprise entre le 01/03/2015 et le 31/08/2021. La DP est fixée au 31/08/2021.

BREADNER et al. [24] ont inclus les données de patients débutant une ligne de traitement avec de l'oxaliplatin + 5-FU/capécitabine entre 2008 et 2011. La DO correspond à la date de début de ligne de traitement pour chaque patient, comprise entre le 01/01/2008 et le 31/12/2011. La DDN varie pour chaque patient et est également comprise entre le 01/01/2008 et le 31/12/2011. La DP est fixée au 31/12/2011.

3.2.1.3 Censure et types de données de survie

Il est fréquent que le temps d'un événement soit inconnu pour certains individus. Ce phénomène, connu sous le nom de censure, survient lorsque la date exacte de survenue d'un événement est indéterminée. Les types de censure comprennent :

- **Censure à gauche** : L'événement s'est produit avant la date d'origine, de sorte que le temps exact de l'événement est inconnu, mais il est inférieur à un certain temps. Dans le contexte des analyses de SSP chez les patients atteints de cancer, où la DO correspond à la date de début de ligne de traitement, la censure à gauche est rare. Cela est dû au fait que la décision de commencer un traitement est basée sur une évaluation clinique préalable de l'état du patient. Pour les analyses de SG, la censure

à gauche n'a pas de sens, car un traitement n'est pas commencé après le décès du patient.

- **Censure par intervalle** : L'événement se produit entre deux points d'observation. Par exemple, si un patient est vu à des visites régulières, l'événement est survenu entre deux visites successives.

Dans le contexte des analyses de SSP chez les patients atteints de cancer, où la DO correspond à la date de début de ligne de traitement, bien que théoriquement possible, la censure par intervalle est peu courante. En effet, la progression tumorale est généralement détectée lors des visites régulières et bien documentées. Pour les analyses de SG, la censure par intervalle n'a également pas de sens, car il est impossible pour un patient décédé d'être observé vivant lors de visites ultérieures.

- **Censure à droite** : L'événement n'a pas été observé pendant la période d'étude (avant la DP). Ce type de censure est le plus commun, en particulier dans les études de survie chez les patients atteints de cancer où la DO correspond à la date de début de ligne de traitement. La censure à droite signifie que le patient est encore en vie ou n'a pas encore rechuté à la DP. Ce type de censure peut se présenter sous deux formes :

- **Exclu vivant** : Le patient est encore en vie ou bien l'événement n'a pas eu lieu à la DP, et donc le temps d'événement est inconnu et supérieur à cette date. La date conservée correspond alors à la DP.
- **Perdu de vue** : L'information sur le patient est incomplète, et on ne sait pas si l'événement a eu lieu ou non après le dernier point d'observation connu. Les patients perdus de vue peuvent l'être par exemple en raison d'un retrait de consentement à l'étude, ou d'un déménagement. La date conservée correspond alors à la DDN.

Dans les études de MENG et al. [131] et BREADNER et al. [24], les cas de censure sont principalement à droite, où les patients n'ont pas encore vécu l'événement d'intérêt à la date de fin de l'étude ou ont été perdus de vue. Les censures à gauche et par intervalle sont rares pour les analyses de SSP et n'ont pas de sens pour les analyses de SG, étant donné les définitions des événements d'intérêt et les contextes cliniques.

3.2.1.4 Le temps de survie, une variable continue

Le temps, le délai ou la durée de survie correspond au temps écoulé jusqu'à l'apparition d'un événement spécifique depuis la date d'origine.

Soit C_i le délai de survenue de la censure et T_i^* le délai de survenue de l'événement d'intérêt pour un individu i . La fonction indicatrice de l'événement, notée d_i , est définie par :

$$d_i = I(T_i^* \leq C_i) \quad (3.1)$$

où I est la fonction indicatrice.

Alors pour un individu i :

- si $T_i^* \leq C_i$, cela signifie que le temps de survenue de l'événement est connu et on note $d_i = 1$;
- si $T_i^* > C_i$, le temps de survenue de l'événement n'est pas connu, ou bien l'événement n'a pas eu lieu, et on note $d_i = 0$.

Soit T_i le temps de survie observé, T_i est une variable aléatoire définie par $T_i = \min(T_i^*, C_i)$, où T_i^* et C_i sont respectivement les variables aléatoires associées aux temps de survenue de l'événement et de censure. La variable aléatoire T_i est non négative et sa distribution est supposée continue.

3.2.1.5 Le format des données en survie

En analyse de survie, les données observées sont représentées par l'ensemble $D_n = \{(T_i, d_i, Z_i); i = 1, \dots, n\}$ [195], où :

- T_i est le temps observé pour le i -ème individu, défini comme $T_i = \min(T_i^*, C_i)$;
- $d_i = I(T_i \leq C_i)$ est l'indicateur de l'événement, avec $d_i = 1$ si l'événement est observé et $d_i = 0$ si les données sont censurées;
- $Z_i \equiv (Z_{i1}, \dots, Z_{iJ})$ est le vecteur des covariables observées pour l'individu i .

Par exemple, le format des données pourrait inclure des covariables telles que l'âge, le sexe, et d'autres caractéristiques cliniques, comme illustré ci-dessous :

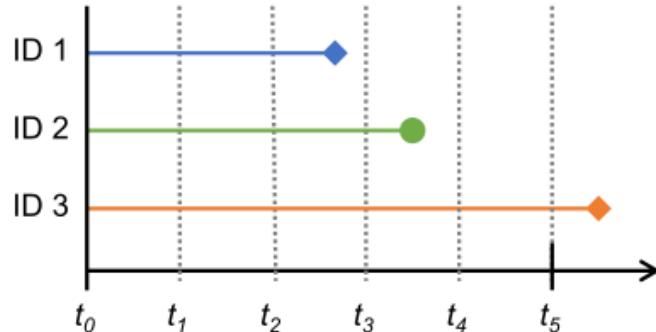


FIGURE 3.4 – Représentation graphique des timelines de patients 1, 2 et 3. Les cercles indiquent une censure tandis que les losanges indiquent la survenue de l'événement. Pour le patient 1, l'événement est survenu à $T_1 = 2.7$ et $d_1 = 1$ indique que l'événement a été observé. Pour le patient 2, le temps observé est $T_2 = 3.4$ et $d_2 = 0$ indique que les données sont censurées à cette date. Pour le patient 3, l'événement est survenu à $T_3 = 5.6$ et $d_3 = 1$ indique que l'événement a été observé. L'image a été adaptée et provient de l'article de SURESH, SEVERN et GHOSH [195].

ID	Temps (T_i)	Événement (d_i)	Âge (Z_{i1})	Sexe (Z_{i2})	Autres covariables (Z_{i3}, \dots, Z_{iJ})
1	2.7	1	45	M	...
2	3.4	0	50	F	...
3	5.6	1	60	M	...

TABLE 3.1 – Format des données de survie avec covariables pour les patients 1, 2 et 3 représentés dans la figure 3.4. Les covariables peuvent inclure des caractéristiques cliniques comme l’âge, le sexe, et d’autres informations spécifiques aux patients. Par exemple, dans l’étude de MENG et al. [131], les covariables incluent l’âge, le sexe, la présence de métastases, la chirurgie de la tumeur primitive, et le nombre de sites métastatiques. Dans l’étude de BREADNER et al. [24], les covariables incluent l’âge, le sexe, la dose-intensité relative (RDI), le stade tumoral et le stade nodal.

3.2.1.6 Fonction de survie et fonction de risque instantané

Plusieurs fonctions clés sont utilisées en analyse de survie.

Définition - Fonction de répartition F

Soit T le temps de survie. Si T est une variable continue, on peut définir sa fonction de densité $f(t)$ (ou densité de probabilité). Sa fonction de répartition $F(t)$ est alors définie comme

$$F(t) = P(T \leq t) = \int_0^t f(u) \, du \quad (3.2)$$

Définition - Fonction de survie S

La fonction de survie S est la probabilité de ne pas avoir subi l’événement d’intérêt avant le temps t , soit la proportion des sujets n’ayant pas fait l’évènement juste avant le temps t . Elle se définit comme

$$S(t) = 1 - F(t) = P(T > t). \quad (3.3)$$

$S(0) = 1$ indique que la probabilité de survie est totale au début du suivi, c’est-à-dire que tous les patients sont vivants, et $S(\infty) = 0$ indique que tous les patients vivront l’événement à un moment donné.

Définition - Fonction de risque instantané h

La fonction de risque instantané h est la probabilité de survenue de l’événement d’intérêt dans un petit intervalle de temps juste après t , étant donné que l’individu a survécu

jusqu'à ce moment t :

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \quad (3.4)$$

Notons que la fonction de survie $S(t)$ porte sur l'absence d'événement tandis que $h(t)$ porte sur sa survenue.

Définition - Fonction de risque cumulé H

La fonction de risque cumulé est une quantité intégrale qui représente la somme des risques instantanés dans le temps, et est définie comme :

$$H(t) = \int_0^t h(u) du \quad (3.5)$$

Cette fonction donne une mesure globale du risque de l'événement survenant jusqu'au temps t .

Propriété - Lien entre les fonctions f , F , H et h

Les fonctions $f(t)$, $S(t)$, $H(t)$ et $h(t)$ sont liées par les relations suivantes :

$$\begin{cases} h(t) = \frac{f(t)}{S(t)} \\ S(t) = \exp(-H(t)) = \exp\left(-\int_0^t h(u) du\right) \end{cases} \quad (3.6)$$

Dans l'analyse de survie classique, on cherche à estimer les fonctions $S(t)$ (*cf. section 3.2.1.7*) et $h(t)$ (*cf. section 3.2.1.10*) avec les données de survie (*cf. section 3.2.1.5*).

3.2.1.7 L'estimation non paramétrique de Kaplan-Meier (KM)

L'estimation de Kaplan-Meier (KM) est une méthode non paramétrique très fréquemment utilisée pour estimer la fonction de survie [108]. Elle produit une courbe de survie en escalier, où chaque palier représente une estimation de la probabilité de survie jusqu'à un certain temps :

$$\hat{S}_{KM}(t) = \prod_{k:t_k \leq t} \left(1 - \frac{\nu_k}{n_k}\right) \quad (3.7)$$

où t_k sont les temps où au moins un événement a eu lieu, ν_k est le nombre d'événements à t_k , et n_k est le nombre de sujets à risque juste avant t_k .

Les courbes de Kaplan-Meier pour les deux études illustrent visuellement les différences de survie entre les groupes définis par les variables d'intérêt (*cf. figure 3.5*).

3.2.1.8 Le test du log-rank

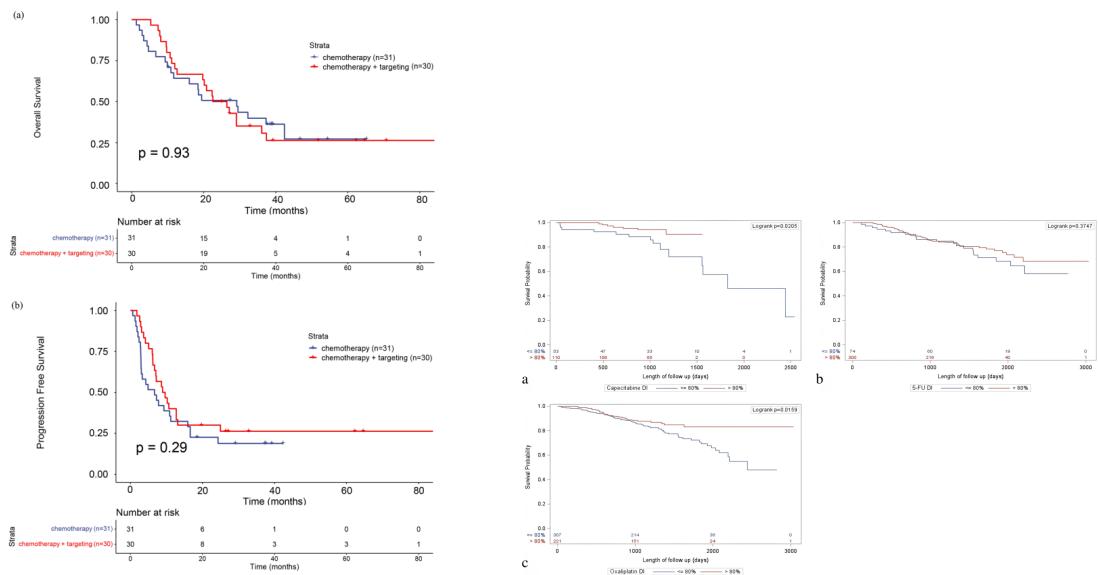
Le test du log-rank est un outil statistique non paramétrique pour la comparaison des fonctions de survie entre deux ou plusieurs groupes comme des bras dans une étude clinique

[125]. Il se base sur la comparaison des temps d'événement observés avec ceux attendus sous l'hypothèse nulle d'égalité des fonctions de survie. La statistique du test du log-rank s'écrit :

$$U_{log-rank} = \sum_{j=1}^J \frac{(O_j - E_j)^2}{E_j} \quad (3.8)$$

où O_j est le nombre d'événements observés et E_j est le nombre d'événements attendus dans le groupe j , et J est le nombre de groupes. La p-valeur est obtenue en référence à une distribution du χ^2 avec $J - 1$ degrés de liberté [124]. La p-valeur obtenue indique la significativité statistique de la différence entre les courbes de survie. Lorsque plus de deux groupes sont testés, une p-valeur significative signifie qu'au moins un des groupes a une courbe de survie différente des autres.

Dans l'étude de MENG et al. [131], les p-valeurs des tests de log-rank montrent que les différences de survie entre les patients recevant une chimiothérapie seule versus ceux recevant une chimiothérapie avec thérapie ciblée ne sont pas statistiquement significatives pour la SG ($p = 0.93$) et la SSP ($p = 0.29$) (cf. figure 3.5). Dans celle de BREADNER et al. [24], les p-valeurs des test de log-rank sont significatives pour la capécitabine ($p = 0.0205$) et pour l'oxaliplatine ($p = 0.0159$), mais pas pour le 5-FU ($p = 0.3747$), indiquant une meilleure survie pour les groupes avec une RDI > 80% pour la capécitabine et l'oxaliplatine.



(a) Courbes de Kaplan-Meier de SSP et de SG dans l'étude de MENG et al. [131]

(b) Courbes de Kaplan-Meier de SG dans l'étude de BREADNER et al. [24]

FIGURE 3.5 – Figure 3.5a : Courbes de Kaplan-Meier de la survie sans progression (SSP) et de survie globale (SG) dans l'étude de MENG et al. [131], comparant les patients recevant une chimiothérapie seule versus ceux recevant une chimiothérapie avec thérapie ciblée. Les p-valeurs des tests de log-rank montrent que les différences ne sont pas statistiquement significatives pour la SG ($p = 0.93$) et la SSP ($p = 0.29$). Graphiquement, cela est visible car les courbes se croisent.

Figure 3.5b : Courbes de Kaplan-Meier de la survie globale (SG) dans l'étude de BREADNER et al. [24], comparant les patients avec une $\text{RDI} \leq 80\%$ et ceux avec une $\text{RDI} > 80\%$. Les p-valeurs des tests de log-rank montrent une différence statistiquement significative pour la capécitabine ($p = 0.0205$) et l'oxaliplatine ($p = 0.0159$), mais pas pour le 5-FU ($p = 0.3747$). Graphiquement, on observe en effet que les courbes se croisent dans le graphique de l'oxaliplatine, ce qui n'est pas le cas des deux autres graphiques.

L'estimation de Kaplan-Meier et le test du log-rank ne permettent pas de quantifier l'effet de covariables explicatives ni d'ajuster un modèle sur ces covariables.

3.2.1.9 La modélisation semi-paramétrique de Cox

Définition - Modèle de Cox

Le modèle de Cox à risques proportionnels est un modèle semi-paramétrique permettant d'évaluer l'impact de plusieurs variables explicatives sur le risque de survenue de l'événement d'intérêt [44]. Soient Z_1, Z_2, \dots, Z_p des covariables explicatives, le modèle s'écrit :

$$h(t) = h_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2 + \dots + \beta_p Z_p) = h_0(t) \exp\left(\sum_{j=1}^p \beta_j Z_j\right) \quad (3.9)$$

où $h_0(t)$ est le risque de base, et les β_i les coefficients à estimer.

Dans l'étude de MENG et al. [131], dans un premier temps, des modèles univariés de Cox sont utilisés pour évaluer l'impact de 14 variables explicatives sur le risque de progression tumorale (SSP) ou de décès (SG). La variable d'intérêt, le type de traitement encodé de façon binaire (chimiothérapie avec ou sans bevacizumab), fait partie des 14 analyses univariées. Les autres variables incluent par exemple l'âge (binaire, ≤ 65 ans vs. > 65 ans), le sexe (binaire, F/M), la présence de métastases dans le foie (binaire, oui/non), la chirurgie de la tumeur primitive (binaire, oui/non) et le nombre de sites métastatiques (binaire, unique/multiple). À titre d'exemples, les modèles univariés pour le type de traitement, l'âge et le sexe s'écrivent de cette façon :

$$h(t) = h_0(t) \exp(\beta_1 \text{type de traitement de première ligne})$$

$$h(t) = h_0(t) \exp(\beta_1 \text{âge})$$

$$h(t) = h_0(t) \exp(\beta_1 \text{sexe})$$

Une fois que les modèles univariés sont ajustés, les contributions des coefficients β sont estimées (*cf.* section 3.2.1.10), et un modèle multivarié est ajusté sur les variables significatives des modèles univariés. Trois variables se sont révélées significatives pour la SG (présence de métastases dans le foie, chirurgie de la tumeur primitive et le nombre de sites métastatiques) dans les analyses univariées, et deux variables pour la SSP (chirurgie de la tumeur primitive et la présence de métastases dans le foie). Par exemple, pour la SSP, le modèle multivarié s'écrit :

$$h(t) = h_0(t) \exp(\beta_1 \text{présence de métastases dans le foie} + \beta_2 \text{chirurgie de la tumeur primitive})$$

Dans l'étude de BREADNER et al. [24], trois analyses multivariées sont directement réalisées pour évaluer l'impact de la RDI des chimiothérapies (oxaliplatin + 5-FU/capécitabine) sur la SG. Les variables incluent l'âge, le sexe, la RDI ($\leq 80\%$ vs. $> 80\%$), le stade tumoral (T0-T2 vs. T3 vs. T4) et le stade nodal (N0 vs. N1 vs. N2) (*cf.* TNM section 1.2.1). Les modèles pour l'oxaliplatin, le 5-FU, la capécitabine s'écrivent de cette façon :

$$h(t) = h_0(t) \exp(\beta_1 \text{RDI d'oxaliplatin} + \beta_2 \text{âge} + \beta_3 \text{sexe} + \beta_4 \text{stade tumoral} + \beta_5 \text{stade nodal})$$

$$h(t) = h_0(t) \exp(\beta_1 \text{RDI de 5-FU} + \beta_2 \text{âge} + \beta_3 \text{sexe} + \beta_4 \text{stade tumoral} + \beta_5 \text{stade nodal})$$

$$h(t) = h_0(t) \exp(\beta_1 \text{RDI de capécitabine} + \beta_2 \text{âge} + \beta_3 \text{sexe} + \beta_4 \text{stade nodal})$$

3.2.1.10 Estimation des coefficients dans le modèle de Cox et hazard ratios

Pour estimer les coefficients β dans le modèle de Cox, nous faisons l'hypothèse que les temps de survenue des événements sont indépendants et identiquement distribués (i.i.d) conditionnellement aux covariables. Cette hypothèse signifie que, bien que les individus aient des risques différents basés sur leurs covariables, ces risques sont indépendants les uns des autres. Cela nous permet d'exprimer la vraisemblance comme le produit des probabilités individuelles.

L'estimation des coefficients β est obtenue par la maximisation de la vraisemblance partielle. La vraisemblance, notée $L(\beta)$, est une fonction qui mesure à quel point les valeurs des coefficients β rendent les données observées probables. La vraisemblance partielle se concentre spécifiquement sur les temps où des événements ont été observés. Elle permet de se concentrer uniquement sur les paramètres β , en intégrant l'effet des covariables sans nécessiter de spécification du risque de base $h_0(t)$.

Définition - Fonction de vraisemblance partielle L_p

La fonction de vraisemblance partielle $L_p(\beta)$ pour le modèle de Cox est définie comme

$$L_p(\beta) = \prod_{i=1}^n \left(\frac{\exp\left(\sum_{j=1}^p \beta_j Z_{ij}\right)}{\sum_{l \in R_{Cox}(T_i)} \exp\left(\sum_{j=1}^p \beta_j Z_{lj}\right)} \right)^{d_i} \quad (3.10)$$

où

- T_i est le temps observé pour l'individu i , défini comme $T_i = \min(T_i^*, C_i)$.
- d_i est l'indicateur de l'événement pour l'individu i , avec $d_i = 1$ si l'événement est observé et $d_i = 0$ si les données sont censurées.
- $Z_i \equiv (Z_{i1}, \dots, Z_{iJ})$ est le vecteur des covariables observées pour l'individu i .
- $R_{Cox}(T_i)$ est l'ensemble des sujets à risque au temps T_i .

Cette formule exprime que la vraisemblance partielle se concentre sur les risques relatifs $\exp\left(\sum_{j=1}^p \beta_j Z_{ij}\right)$ comparés aux risques totaux $\sum_{l \in R_{Cox}(T_i)} \exp\left(\sum_{j=1}^p \beta_j Z_{lj}\right)$ au temps T_i . La fonction indicatrice d_i assure que seuls les individus ayant eu l'événement contribuent directement à la vraisemblance.

Pour obtenir les coefficients β , on maximise cette fonction de vraisemblance partielle. Cela permet de trouver les valeurs des coefficients β qui rendent les données observées les plus probables.

L'estimation des coefficients β permet de quantifier la contribution de chacune des variables explicatives au modèle. Une contribution positive indique une augmentation du risque, tandis qu'une contribution négative indique une diminution du risque. La significativité de ces contributions est testée avec le test de Wald [90].

Bien que l'estimation des coefficients β et de leurs p-valeurs apporte une réponse statistique valide sur l'impact des co-variables explicatives, cette estimation est rarement présentée dans les études cliniques. En effet, dans ces études, les résultats des régressions de Cox sont généralement interprétés avec des hazard ratios (HR). Le hazard ratio est le rapport des risques instantanés de survenue de l'événement entre deux groupes différents. Un HR supérieur à 1 indique un risque accru d'événement dans le groupe de référence par rapport au groupe de comparaison, tandis qu'un HR inférieur à 1 indique un risque réduit.

Définition - Hazard Ratio *HR*

La formule pour calculer le hazard ratio est la suivante :

$$HR = \frac{h(t|Z_j = a)}{h(t|Z_j = b)} = \exp(\beta_j(a - b)) \quad (3.11)$$

où $h(t|Z_j = a)$ et $h(t|Z_j = b)$ représentent les risques instantanés pour les groupes avec les valeurs a et b de la variable Z_j , respectivement.

Dans l'étude de MENG et al. [131], le hazard ratio pour mesurer l'impact du type de traitement (chimiothérapie avec ou sans bérvacizumab) sur la SG s'écrit comme suit :

$$HR_{SG} = \frac{h(t|\text{chimiothérapie + bérvacizumab})}{h(t|\text{chimiothérapie seule})} = 1.027$$

Dans les analyses univariées avec la variable type de traitement (*cf. figure 3.6 colonne “Univariable analysis”*), le hazard ratio pour la SG est de 1.027, et pour la SSP est de 0.734, mais les p-valeurs des coefficients β associées sont supérieures à 0.05. L'ajout de bérvacizumab n'améliore pas significativement le SG ni la SSP par rapport à la chimiothérapie seule. En revanche, les analyses univariées ont montré que la présence de métastases dans le foie, la chirurgie de la tumeur primitive et le nombre de sites métastatiques étaient significativement associés à la SG, tandis que la chirurgie de la tumeur primitive et la présence de métastases dans le foie étaient significativement associés à la SSP. Les analyses multivariées (*cf. figure 3.6 colonne “Multivariable analysis”*) ont confirmé ces résultats, avec des hazard ratios de 2.399 pour les métastases dans le foie (SG) et de 0.523 pour la chirurgie de la tumeur primitive (SSP), indiquant une augmentation significative du risque pour les patients avec métastases et une diminution significative du risque pour ceux ayant subi une chirurgie de la tumeur primitive.

Dans l'étude de BREADNER et al. [24], les analyses multivariées (*cf. figure 3.7*) ont montré la RDI > 80% était significativement associée à une meilleure SG pour les patients traités avec du 5-FU (HR = 0.56, p = 0.03) et capécitabine (HR = 0.23, p = 0.006), ainsi que pour tous les patients traités avec oxaliplatin (HR = 0.52, p = 0.005). Par exemple, pour les patients traités avec du 5-FU, le hazard ratio s'écrit :

$$HR_{5-FU} = \frac{h(t|RDI > 80\%)}{h(t|RDI \leq 80\%)} = 0.56$$

Ces résultats montrent, pour les trois molécules anticancéreuses, une diminution significative du risque de décès pour les patients ayant une RDI > 80%.

Table 3 Univariate and multivariate analysis of OS

OS	Univariable analysis		Multivariable analysis		
	Characteristic	HR (95%CI)	P-value	HR (95%CI)	P-value
Age(</≥ 65 years)	0.882(0.457–1.705)	0.710			
Gender(Male/Female)	1.108(0.600–2.046)	0.743			
EcoG(< 2/≥ 2)	0.947(0.368–2.437)	0.909			
Tumor site(Multi-sides/Right colon /Left colon/Rectum)	0.920(0.666–1.271)	0.613			
Histologic grade(Low grade/High grade)	0.954(0.486–1.873)	0.891			
Number of metastatic sites(Single/Multiple)	2.003(1.071–3.745)	0.030*	1.603(0.842–3.051)	0.151	
MMR Status(dMMR/pMMR)	1.367(0.323–5.782)	0.671			
Primary tumor surgery(No/Yes)	0.285(0.152–0.537)	< 0.001*	0.326(0.169–0.631)	0.001*	
Intestinal obstruction(No/Yes)	0.655(0.233–1.841)	0.422			
Liver metastases(No/Yes)	2.480(1.302–4.724)	0.006*	2.399(1.242–4.635)	0.009*	
Lung metastases(No/Yes)	0.839(0.427–1.648)	0.610			
Peritoneal metastasis(No/Yes)	1.162(0.536–2.518)	0.704			
Distant lymph node metastases(No/Yes)	0.907(0.462–1.782)	0.778			
First-line medication(Chemo/Chemo + targeted therapy)	1.027(0.555–1.901)	0.932			

Table 4 Univariate and multivariate analysis of PFS

PFS	Univariable analysis		Multivariable analysis		
	Characteristic	HR (95%CI)	P-value	HR (95%CI)	P-value
Age(</≥ 65 years)	0.561(0.295–1.064)	0.077			
Gender(Male/Female)	1.192(0.670–2.121)	0.551			
EcoG(< 2/≥ 2)	1.005(0.397–2.546)	0.991			
Tumor site(Multi-sides/Right colon /Left colon/Rectum)	0.941(0.692–1.281)	0.700			
Histologic grade(Low grade/High grade)	1.350(0.718–2.537)	0.351			
Number of metastatic sites(Single/Multiple)	1.445(0.814–2.564)	0.209			
MMR Status(dMMR/pMMR)	4.070(0.553–29.943)	0.168			
Primary tumor surgery(No/Yes)	0.438(0.243–0.790)	0.006*	0.523(0.280–0.974)	0.041*	
Intestinal obstruction(No/Yes)	1.657(0.739–3.716)	0.221			
Liver metastases(No/Yes)	2.058(1.132–3.742)	0.018*	1.677(0.890–3.157)	0.109	
Lung metastases(No/Yes)	0.667(0.355–1.254)	0.208			
Peritoneal metastasis(No/Yes)	1.813(0.840–3.915)	0.130			
Distant lymph node metastases(No/Yes)	0.702(0.364–1.353)	0.291			
First-line medication(Chemo/Chemo + targeted therapy)	0.734(0.413–1.304)	0.291			

FIGURE 3.6 – Résultats des analyses de régression de Cox : analyses univariées et multivariées des facteurs de risque pour la SG et la SSP dans l'étude de MENG et al. [131].

3.2.1.11 Hypothèses du modèle de Cox

Le modèle de Cox repose sur les hypothèses suivantes :

- Le risque de base $h_0(t)$ est non-paramétrique.
- Dans l'étude de MENG et al. [131], le risque de base de survenue de l'événement

Table 2 Multivariable analyses of associations between patient characteristics and overall survival

Characteristic	Category	FOLFOX (5-FU dose intensity) (n = 371)		CAPOX (Capecitabine dose intensity) (n = 164)		All patients (Oxaliplatin dose intensity) (n = 526)	
		Hazard Ratio (95% Confidence Interval)	P value	Hazard Ratio (95% Confidence Interval)	P value	Hazard Ratio (95% Confidence Interval)	P value
Age at chemo start		0.99 (0.97 to 1.02)	0.62	1.01 (0.96 to 1.06)	0.70	1.00 (0.98 to 1.02)	0.95
Sex	Female	Reference	0.59	Reference	0.18	Reference	0.46
	Male	1.13 (0.71 to 1.80)		2.05 (0.73 to 5.77)		1.17 (0.77 to 1.77)	
Dose intensity	≤80%	Reference	0.03	Reference	0.006	Reference	0.005
	>80%	0.56 (0.33 to 0.94)		0.23 (0.08 to 0.65)		0.52 (0.33 to 0.82)	
T stage	T0-T2	Reference	0.03	–	–	Reference	<0.001
	T3	1.47 (0.58 to 3.75)		–	–	1.66 (0.65 to 4.19)	
	T4	2.56 (1 to 6.6)		–	–	3.33 (1.32 to 8.41)	
N stage	N0	Reference	<0.001	Reference	0.03	Reference	<0.001
	N1	7.20 (0.96 to 53.72)		1.44 (0.16 to 13.05)		7.55 (1.03 to 55.51)	
	N2	20.70 (2.79 to 153.3)		4.70 (0.57 to 38.85)		19.72 (2.71 to 143.54)	

FIGURE 3.7 – Résultats des analyses de régression de Cox : analyses multivariées des facteurs de risque pour la SG dans l'étude de BREADNER et al. [24].

(progression ou décès) est déterminé par les données, sans supposer de forme spécifique.

- Les effets des covariables sont additifs et linéaires.
 - Dans l'étude de BREADNER et al. [24], les variables explicatives telles que l'âge, le sexe et la RDI sont supposées avoir des effets indépendants et linéaires sur la survenue de l'événement.
- Les risques sont proportionnels.
 - Dans l'étude de MENG et al. [131], un hazard ratio de 2.399 pour la présence de métastases dans le foie signifie que le risque relatif de décès pour les patients avec métastases est 2.399 fois plus élevé que pour ceux sans métastases, et ce rapport reste constant dans le temps.
 - Le temps t est automatiquement ajusté.
 - Dans les deux études, les patients sont suivis pendant des durées variables. Le modèle de Cox ajuste automatiquement ces différences de suivi sans transformation particulière des données de temps.

Le modèle de Cox, dit à risques proportionnels (*proportional hazards*), suppose que le rapport des risques (*hazard ratio*) entre deux individus reste constant dans le temps, ce qui est connu sous le nom d'hypothèse des risques proportionnels. On a :

$$\frac{h(t|Z_1, \dots, Z_j, \dots, Z_p)}{h(t|Z_1, \dots, 0, \dots, Z_p)} = \exp(\beta_j Z_j)$$

soit le taux est constant au cours du temps. Cette hypothèse doit être vérifiée à l'aide de tests statistiques appropriés, tels que le test basé sur les résidus de Schoenfeld [179].

3.2.1.12 Distribution et modèles paramétriques

Le modèle de Cox ne suppose pas de distribution spécifique de la fonction de risque $h(t)$. Il existe aussi des modèles paramétriques pour modéliser la fonction $h(t)$. Ceux-ci

offrent une alternative utile lorsque l'on dispose d'informations a priori sur la distribution du temps jusqu'à l'événement. Par exemple, le modèle exponentiel suppose une fonction de risque constante dans le temps [25], tandis que le modèle de Weibull permet des fonctions de risque croissantes, constantes ou décroissantes [29], et est souvent utilisé pour modéliser l'obsolescence programmée des appareils électroniques. Toutefois, nous ne détaillerons pas ces modèles car nous ne les avons pas utilisés au cours de cette thèse.

Bien que la régression de Cox soit massivement utilisée en médecine pour étudier les survenues d'événements de santé, la régression logistique, peut aussi être utilisée pour modéliser la survenue d'événements.

3.2.2 Régression logistique pour l'analyse de survie

Dans son étude, ABBOTT [1] examine l'impact de l'âge, l'indice de masse corporelle, le tabagisme et le taux de cholestérol sur la survenue de maladies coronariennes chez des hommes âgés de 40 à 49 ans. Il modélise cette survenue avec des régressions de Cox et logistique, et compare la significativité des contributions de chaque facteur de risque des deux modèles. Les résultats des deux modèles convergent : le tabagisme et l'indice de masse corporelle sont significativement corrélés à la survenue de maladie coronarienne, tandis que le taux de cholestérol et l'âge ne le sont pas. Il conclut ainsi que la régression logistique peut être une méthode valide pour l'analyse de survie, particulièrement dans les cas où la date exacte de la survenue de l'événement n'est pas identifiée, mais qu'on sait qu'elle est survenue dans un intervalle de temps. La régression logistique est parfois utilisée en oncologie pour modéliser la probabilité qu'un événement comme la progression, la rechute du cancer ou le décès survienne [149], mais cela reste assez marginal.

Rappelons que le modèle de Cox est basé sur un modèle de temps continu (*cf.* 3.2.1.4). Dans cette section, nous expliquons les bases de la régression logistique et son application à l'analyse de survie, en distinguant les modèles de temps continu [223] et discret [1, 195].

3.2.2.1 Modèle de temps continu

Dans un modèle de temps continu, les temps de survenue des événements sont mesurés avec précision et peuvent prendre n'importe quelle valeur sur une échelle continue. Nous avons vu dans les sections 3.2.1.4, 3.2.1.5 et 3.2.1.6 que ce modèle de temps continu était adopté pour l'analyse de survie classique.

Modélisation de la régression logistique

Dans le modèle de Cox, l'objectif est d'estimer la fonction de risque instantané $h(t)$, qui représente le taux de survenue de l'événement à un instant t donné, conditionnellement au fait que l'individu ait survécu jusqu'à cet instant. En régression logistique appliquée à la survie, nous cherchons à estimer la fonction logit de la probabilité que l'événement survienne à un temps donné t_j . Cette probabilité est modélisée par une variable binaire Y_i qui indique, pour l'individu i , la survenue (1) ou non (0) de l'événement à ce temps t_j .

Soit Y la variable binaire qui indique la survenue de l'événement, avec $Y_i = 1$ si l'événement survient pour un individu i à un point de temps spécifique t_j , et $Y_i = 0$ sinon. Les

facteurs de risque sont représentés par X_1, X_2, \dots, X_r . Le modèle de régression logistique peut être écrit comme suit :

$$\text{logit}(P(Y_i = 1)) = \log \left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)} \right) = \alpha + \sum_{k=1}^r \beta_k X_k \quad (3.12)$$

où $P(Y_i = 1)$ est la probabilité que l'événement survienne pour l'individu i , α est l'intercept, une constante, et β_k sont les coefficients des facteurs de risque X_k .

Estimation des coefficients β

Les coefficients β sont estimés en maximisant la vraisemblance, donnée par :

$$L(\beta) = \prod_{i=1}^n P(Y_{ij} = 1)^{d_{ij}} (1 - P(Y_{ij} = 1))^{1-d_{ij}} \quad (3.13)$$

où d_{ij} est l'indicateur de l'événement pour l'individu i au temps t_j , avec $d_{ij} = 1$ si l'événement est observé au temps t_j et $d_{ij} = 0$ sinon.

L'estimation des coefficients β permet de quantifier leur contribution au modèle. Cependant, l'interprétation des coefficients β dans un modèle de régression logistique est moins directe que dans un modèle de Cox. Une contribution positive indique une augmentation des log-odds de l'événement, tandis qu'une contribution négative indique une diminution des log-odds de l'événement. Concrètement, cela signifie qu'une variable avec un coefficient positif augmente la probabilité de survenue de l'événement, tandis qu'une variable avec un coefficient négatif la diminue. La significativité de ces contributions est testée avec le test de Wald [90].

Dans l'étude de ZABOR et al. [223], les résultats des régressions logistiques univariées indiquent que seul le stade tumoral est significativement corrélé à la survenue de toxicité de grade supérieur à 3. Cela signifie que le stade tumoral augmente significativement la probabilité de survenue de cette toxicité.

Format des données

Le format des données de survie dans un modèle continu est analogue à celui décrit dans la Table 3.1 dans l'analyse de survie classique, avec une entrée par patient (*cf.* Figure 3.8).

3.2.2.2 Modèle de temps discret

Dans un modèle de temps discret, le temps est divisé en intervalles réguliers, et les événements sont enregistrés selon ces intervalles. SURESH, SEVERN et GHOSH [195] détaillent l'application de la régression logistique pour les modèles de temps discret. Dans le modèle discret, le temps est divisé en une séquence de J intervalles de temps contigus $(t_0, t_1], (t_1, t_2], \dots, (t_{J-1}, t_J]$, où $t_0 = 0$.

Fonction de survie et probabilité conditionnelle dans le modèle de temps discret

Dans ce modèle, le risque de survenu de l'événement pour un individu i dans un inter-

valle $A_j = (t_{j-1}, t_j]$ est la probabilité pour i de vivre l'événement dans cet intervalle sachant qu'il a vécu jusqu'à cet intervalle. Ainsi le risque $h(t)$ (cf. définition 3.4) défini comme un taux dans le modèle continu est une probabilité conditionnelle dans le modèle discret. La probabilité de survenue de l'événement dans un intervalle de temps donné est modélisée comme suit :

$$\lambda_{ij}(X_i) = P(T_i \in A_j \mid T_i > t_{j-1}, X_i) = P(t_{j-1} < T_i \leq t_j \mid T_i > t_{j-1}, X_i) \quad (3.14)$$

où $\lambda_{ij}(X_i)$ est la probabilité conditionnelle que l'événement survienne dans l'intervalle $A_j = (t_{j-1}, t_j]$ pour l'individu i , étant donné qu'il a survécu jusqu'au début de cet intervalle.

La fonction discrète de probabilité $\lambda_{ij}(X_i)$ est définie comme suit :

$$f_{ij} = \Pr(T_i \in A_j \mid X_i) = S(t_{j-1} \mid X_i) - S(t_j \mid X_i) \quad (3.15)$$

La probabilité de survivre après un certain t est le produit de probabilités conditionnelles d'avoir survécu sur tous les intervalles antérieurs et incluant $A_j = (t_{j-1}, t_j]$, avec $t_j \leq t$. Ainsi la probabilité de survivre dans le modèle de temps discret est donné par :

$$S(t \mid X_i) = \Pr(T_i > t \mid X_i) = \prod_{j:t_j \leq t} (1 - \lambda_{ij}(X_i)) \quad (3.16)$$

Ceci est analogue à la définition du risque cumulé $H(t)$ (cf. formule 3.5), comme l'intégrale de 0 à t de h .

Modélisation de la régression logistique

La probabilité que l'événement survienne dans un intervalle de temps donné est alors modélisée par une régression logistique :

$$\text{logit}(\lambda_{ij}(X_i)) = \alpha_j + \beta X_i \quad (3.17)$$

où α_j est le logit du risque de base pour l'intervalle $(t_{j-1}, t_j]$, et β décrit l'effet des autres covariables sur le risque de base à l'échelle du logit, similaire à une régression logistique.

Estimation des coefficients β

La fonction de vraisemblance pour le modèle de temps discret est donnée par :

$$L = \prod_{i=1}^n \left(\prod_{j=1}^{J_i} \lambda_{ij}(X_i)^{d_{ij}} (1 - \lambda_{ij}(X_i))^{1-d_{ij}} \right) \quad (3.18)$$

où d_{ij} est l'indicateur de l'événement pour l'individu i dans l'intervalle j , avec $d_{ij} = 1$ si l'événement est observé dans cet intervalle et $d_{ij} = 0$ sinon.

Dans les modèles de temps discret, les coefficients β quantifient l'effet des variables explicatives sur les log-odds de survenue de l'événement dans chaque intervalle de temps. Un

coefficient positif signifie une augmentation des log-odds de l'événement dans cet intervalle de temps, tandis qu'un coefficient négatif indique une diminution. Cela implique que les variables explicatives influencent la probabilité de survenue de l'événement de manière spécifique à chaque intervalle de temps. La significativité des coefficients est testée avec le test de Wald [90].

Dans l'étude de ABBOTT [1], les résultats montrent que le tabagisme et l'indice de masse corporelle sont significativement associés à la survenue de maladies coronariennes. Par exemple, un coefficient de tabagisme positif suggère que fumer augmente la probabilité de développer une maladie coronarienne dans les intervalles de temps considérés.

Format des données

Le format des données dans un modèle discret comporte une entrée distincte pour chaque intervalle de temps et chaque individu. Ainsi, chaque individu contribue une ligne pour chaque intervalle de temps pendant lequel il est encore à risque au début de cet intervalle. Chaque enregistrement inclut l'indicateur d'événement pour cet intervalle, un vecteur de covariables de base X_i , et une variable de facteur identifiant l'intervalle A_j correspondant (*cf.* Figure 3.8).

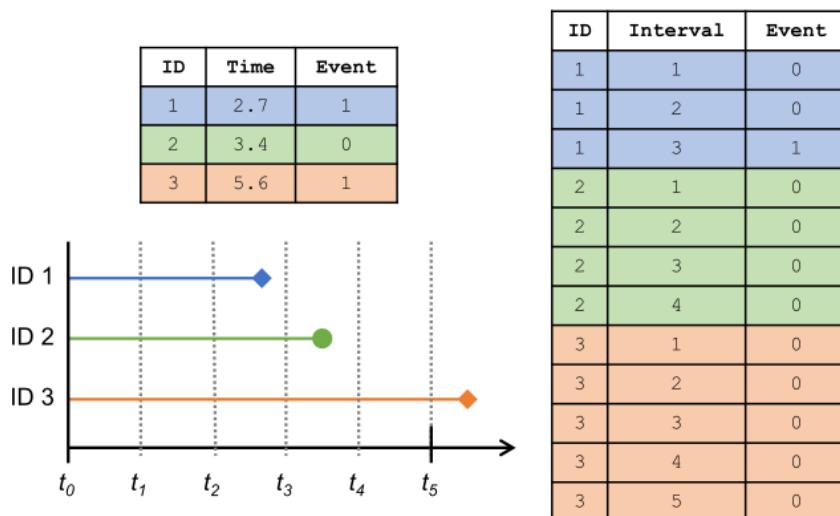


FIGURE 3.8 – Formats des données de survie : temps continu (à gauche) et temps discret (à droite). Le modèle de temps discret utilise le format de droite avec une entrée par intervalle de temps. La figure est tirée de l'article de SURESH, SEVERN et GHOSH [195]

3.2.3 Comparaison des modèles de survie étudiés

3.2.3.1 Modèle de temps continu vs. modèle de temps discret avec une régression logistique

Bien que les modèles de temps continu et discret puissent être utilisés pour l'analyse de

survie avec une régression logistique, le modèle de temps discret présente plusieurs avantages qui le rendent souvent plus approprié dans ce contexte. Comme le souligne ABBOTT [1], les événements en analyse de survie sont fréquemment observés à des intervalles de temps réguliers plutôt qu'à des points spécifiques sur une échelle continue. Par conséquent, modéliser la survenue d'un événement dans un intervalle de temps est plus réaliste et mieux aligné avec la façon dont les données sont collectées dans de nombreuses études cliniques.

De plus, SURESH, SEVERN et GHOSH [195] avancent trois arguments supplémentaires pour préférer le modèle de temps discret en survie, et ce, même à la modélisation continue avec régressions de Cox.

3.2.3.2 Modèle de temps discret avec régression logistique vs. modèle de Cox

Premièrement, les modèles discrets peuvent gérer facilement les temps de survie censurés et les événements concurrents (*competitive risk*) sans nécessiter d'ajustements complexes. Les événements concurrents surviennent lorsque plusieurs types d'événements peuvent empêcher l'observation de l'événement d'intérêt. Par exemple, dans une étude de survie pour des patients atteints de cancer, l'événement d'intérêt pourrait être la récidive du cancer, mais certains patients pourraient décéder avant la récidive, ce qui empêche l'observation de cette dernière. La modélisation en temps discret permet de traiter ces événements en calculant des probabilités conditionnelles pour chaque type d'événement dans chaque intervalle de temps, simplifiant ainsi l'analyse sans avoir besoin de modèles spécifiques pour chaque type d'événement concurrent.

Deuxièmement, la modélisation discrète gère naturellement événements simultanés (*ties*), c'est-à-dire les situations où plusieurs événements surviennent au même moment. Dans les modèles de Cox, les événements simultanés nécessitent des ajustements particuliers comme les méthodes de Breslow [26] ou d'Efron [57] pour estimer correctement les risques. En revanche, dans un modèle de temps discret, les événements simultanés sont directement intégrées dans les probabilités conditionnelles calculées pour chaque intervalle de temps, ce qui simplifie grandement l'analyse.

Enfin, les modèles de temps discret n'exigent pas l'hypothèse des risques proportionnels (cf. section 3.2.1.11). Les modèles discrets permettent d'utiliser n'importe quelle méthode de classification binaire dans le cadre de l'analyse de survie sans nécessiter cette hypothèse.

Cependant, l'interprétabilité directe des résultats des modèles de Cox explique leur large prédominance dans les analyses de survie. Les contributions des variables explicatives à la survenue de l'événement d'intérêt peuvent être directement analysées et quantifiées avec l'estimation de leurs coefficients 3.2.1.10. Cela facilite l'inférence statistique et la communication des résultats par rapport aux résultats d'une régression logistique.

De plus, les modèles de Cox, peuvent être plus efficaces sur le plan computationnel. La modélisation en temps continu ne nécessite pas la transformation des données en intervalles discrets, ce qui peut augmenter considérablement la taille de l'ensemble de données

et la complexité des calculs.

Enfin, les modèles de Cox sont bien établis et largement utilisés dans la recherche clinique. Ils bénéficient d'une riche littérature méthodologique et d'outils logiciels robustes pour l'estimation et la validation des modèles, ce qui peut faciliter leur mise en œuvre et leur adoption dans de nouvelles études.

En résumé, bien que les modèles de temps discret avec classification offrent des avantages pour la gestion des événements concurrents, des événements simultanés, et des hypothèses de risques proportionnels, les modèles de Cox demeurent privilégiés pour leur interprétabilité directe et leur efficacité computationnelle.

3.2.4 Prédictions d'événements avec des modèles de survie

Nous avons vu dans les sections précédentes que l'ajustement de modèles de survie, en particulier ceux de la survie classique, permet de valider ou invalider des hypothèses cliniques sur les facteurs de risques contribuant à la survenue d'un événement. Cependant, au-delà de cette analyse, ces modèles peuvent aussi être utilisés pour faire des prédictions de survie.

Les prédictions sont des estimations probabilistes de la survenue future d'un événement d'intérêt pour un patient donné, basées sur les caractéristiques spécifiques de ce patient. Dans le cadre des modèles de survie, ces prédictions permettent d'estimer le temps restant avant la survenue de l'événement pour des patients individuels, en tenant compte des covariables présentes dans le modèle.

En oncologie, les prédictions de survie sont particulièrement utiles pour développer des nomogrammes. Un nomogramme est un outil graphique qui permet de calculer la probabilité individuelle de survenue d'un événement clinique, comme la rechute ou le décès, en intégrant plusieurs variables pronostiques. Cet outil facilite la prise de décision clinique personnalisée en fournissant des estimations claires et visuelles. Par exemple, VAIDYA et al. [207] ont développé un nomogramme pour évaluer, pour des patients atteints de cancer du poumon non à petites cellules à un stade précoce, les bénéfices individuels d'une chimiothérapie adjuvante. Le nomogramme (QuRNom) est basé sur un score radiomique quantitatif (QuRIS) qui intègre des caractéristiques de texture extraites des images de scanner pour prédire la survie sans récidive. JIANG et al. [106] ont démontré que le score HALP (hémoglobine, albumine, lymphocytes, plaquettes) est un prédicteur indépendant fiable de la survie globale et de la survie sans progression chez des patients atteints de cancer du sein. Ce score a été intégré dans un nomogramme pour fournir des prédictions individualisées.

Dans cette section, nous nous intéressons à la méthodologie utilisée pour construire un nomogramme. Cette méthodologie consiste à ajuster, prédire et évaluer les performances de modèles de survie [12, 95]. Le détail de cette méthodologie nous intéresse car nous nous en inspirons pour optimiser les poids de ProtoDrift (chapitre 5), bien que ProtoDrift ne soit pas un nomogramme.

3.2.4.1 Définir une question clinique, une population de patients et un événement d'intérêt

Pour construire un nomogramme, il est impératif de définir une question clinique claire, une population de patients spécifique et un événement d'intérêt mesurable. Ces éléments assurent la pertinence et l'applicabilité du nomogramme dans un contexte clinique particulier. [12, 95]

Une question clinique bien définie oriente la sélection des variables et la construction du modèle. Les nomogrammes sont utilisés en routine clinique pour aider à la décision cliniques en quantifiant un risque. Il faut donc s'assurer que le risque quantifié par le nomogramme reflète un risque aidant à la prise de décision sur cette question.

Il faut aussi définir clairement la population de patients pour laquelle le nomogramme sera applicable. Cette définition inclut des critères d'inclusion et d'exclusion spécifiques pour assurer que les patients inclus dans l'étude sont homogènes et que les résultats du nomogramme seront généralisables à une population similaire.

Enfin, il faut définir clairement l'événement clinique d'intérêt que le nomogramme vise à prédire. Cela peut inclure des événements tels que la survie sans récidive, la survie sans progression ou la survie globale.

Dans l'étude de VAIDYA et al. [207], la question clinique est de déterminer si les patients atteints de cancer du poumon non à petites cellules à un stade précoce bénéficieront d'une chimiothérapie adjuvante. La population cible comprend des patients diagnostiqués avec ce type de cancer, et l'événement d'intérêt est la survie sans récidive (SSR). Dans celle de JIANG et al. [106], la question clinique porte sur l'évaluation du pronostic chez les patientes atteintes de cancer du sein. La population cible inclut des patientes diagnostiquées avec ce cancer, et les événements d'intérêt sont la survie sans progression (SSP) et la survie globale (SG).

3.2.4.2 Sélectionner des variables explicatives et définir un modèle statistique

Un fois que la question clinique, la population et l'événement d'intérêt établis, il faut identifier des facteurs de risque [12, 95]. Pour les identifier, on peut comme on l'a décrit dans les sections précédentes ajuster des modèles de Cox et logistique et sélectionner les variables explicatives significativement liées à l'événement (*cf.* sections 3.2.1.10, 3.2.2.1.2, 3.2.2.2.3). Lorsque le nombre de variables explicatives est très élevé on peut au préalable utiliser une méthode de sélection de variable telle que la méthode LASSO (*Least Absolute Shrinkage and Selection Operator*) [202].

Il faut ensuite définir un modèle statistique qui explique la survenue de l'événement avec l'ensemble des variables explicatives sélectionnées. Ce modèle peut à nouveau être formulé via une régression de Cox ou logistique.

Il est important de noter que bien que les régressions logistiques soient assez peu uti-

lisées en oncologie dans les études de survie classiques pour étudier l'impact de facteurs de risques sur la survenue d'un événement (*cf.* section 3.2.2), elles sont en revanche beaucoup utilisées, avec un modèle de temps continu, pour la construction de nomogrammes, tant pour la sélection de variables explicatives que pour la définition du modèle à entraîner [93, 20, 76, 114]. À titre indicatif, une recherche PubMed avec les mots clés "nomogram" et "cancer" dans le titre et "logistic regression" dans le titre ou l'abstract mène à 748 résultats.

Pour ces deux étapes, on peut aussi utiliser des modèles d'apprentissage automatique plus sophistiqués tels que les forêts aléatoires appliquées aux données de survie (*Random Survival Forest*) [101] ou des amplifications de gradient (*Survival Gradient Boosting*) [36]. Les modèles d'apprentissage automatique sont particulièrement appréciés lorsque le nombre de variables explicatives est élevé [128, 116].

Dans l'étude de VAIDYA et al. [207], les variables explicatives sont sélectionnées via une méthode de LASSO et des régressions de Cox. Un score de risque nommé QuRIS est calculé via les variables sélectionnées, et le modèle statistique défini est un modèle de Cox multivariés avec QuRIS parmi les variables explicatives. Dans celle de JIANG et al. [106], les variables explicatives sont sélectionnées via des régressions univariées de Cox. Elles sont ensuite introduites dans un modèle multivarié de Cox correspondant au modèle statistique sélectionné.

Quelque soit le modèle statistique sélectionné, il faut ensuite l'entraîner [12, 95].

3.2.4.3 Cycle d'entraînement et de validation du modèle

Le modèle sélectionné dans l'étape précédente comporte des variables explicatives corrélées à la survenue de l'événement. Apriori, bien que ces variables soient corrélées à la survenue de l'événement, elles ne permettent pas de le prédire. L'objectif du cycle d'entraînement et de validation est de rechercher une combinaison optimale des valeurs des coefficients associés aux variables explicatives, afin d'améliorer la prédiction de l'événement.

Le jeu de données de départ (*cf.* table 3.1 et figure 3.8) est divisé en jeux d'entraînement et de validation via des techniques de rééchantillonage telles que la validation croisée [7] ou le bootstrapping [58]. Pour chaque division jeu d'entraînement-jeu de validation, le modèle statistique sélectionné lors de l'étape précédente est ajusté, et ses performances sont évaluées via des métriques adéquates (*cf.* section 3.2.4.4). Ce processus vise d'une part à obtenir un modèle général, c'est-à-dire, qui ne soit pas performant seulement pour une tranche spécifique de la population. Les performances de chaque combinaison de paramètres du modèle correspondent à une moyenne des performances obtenues pour chaque échantillon de validation. De plus, cela permet d'obtenir des intervalles de confiances de ces performances. D'autre part, il vise à ajuster les paramètres du modèle (*tuning parameters*) pour qu'ils aient les meilleurs scores de performance. Sur ce deuxième point, les paramètres optimaux sont sélectionnés :

- soit à la fin du processus, comme ceux ayant mené aux meilleures scores performances du modèle. Dans ce cas, la boucle d'entraînement-validation peut être quali-

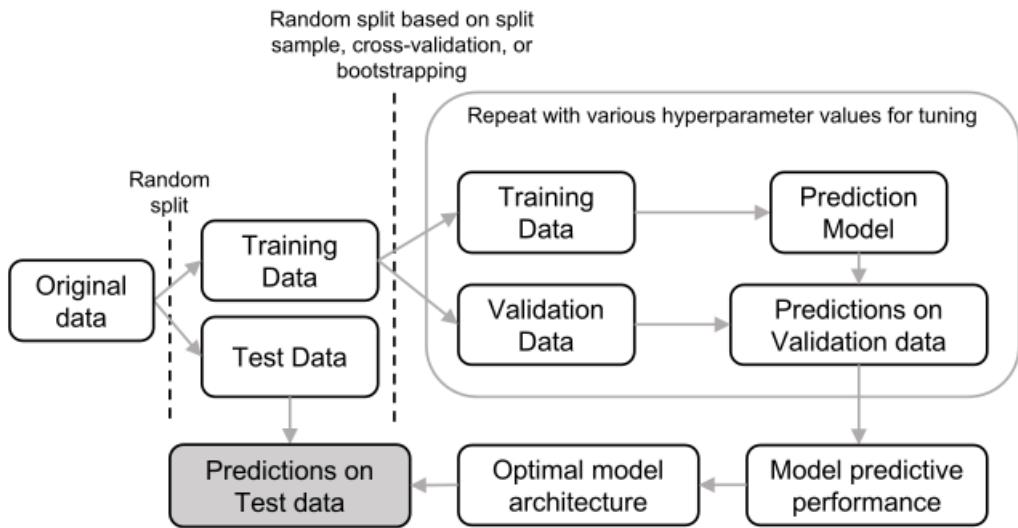


FIGURE 3.9 – Processus de construction et de test du modèle, illustrant la division des données, l’entraînement du modèle et l’évaluation des performances. Figure tirée de l’article de SURESH, SEVERN et GHOSH [195]

fiée d'une recherche par grille (*grid search*),

- soit au cours de la boucle, via des techniques telles que recherche par grille bayésienne (*bayesian grid search*) ou aléatoire (*random grid search*) qui utilisent les résultats de performance de l’itération précédente pour localiser plus efficacement les valeurs de paramètres conduisant aux meilleures performances [18].

Le processus décrit ci-dessus permet d’obtenir le modèle le plus performant sur le jeu de données de départ. Les prédictions finales sont obtenues sur un jeu de données distinct du jeu de données de départ. Selon les disciplines, cette étape est qualifiée de "test du modèle" [195] car on évalue les prédictions sur un jeu de données distinct, appelé le jeu de données test, ou "validation externe" [207, 12] car l'évaluation est identique à celle du cycle entraînement-validation mais sur un jeu de données distinct. Pour les constructions de nomogrammes, il est recommandé que ce jeu de données test provienne d'un hôpital différent du jeu de données d'entraînement.

Dans l'étude de VAIDYA et al. [207] le modèle de Cox incluant QuRIS parmi les variables explicatives est entraîné sur 1 000 échantillons de bootstrap et les scores de prédiction sont obtenus sur deux jeux de données de test provenant d'hôpitaux externes. JIANG et al. [106] n'ont pas entraîné leur modèle. Ils ont ajusté un modèle de Cox multivarié sur leur cohorte principale et évalué la performance des nomogrammes construits à partir de ce modèle en utilisant les mêmes données, sans validation externe ni division explicite des données en jeux d'entraînement et de test.

3.2.4.4 Interpréter les performances du modèle

Dans la section précédente, on a vu que les modèles étaient sélectionnés selon leurs

scores de performance. Il existe divers scores reflétant différentes qualités de performance. Les scores utilisés pour l'optimisation des paramètres du modèle lors de l'entraînement sont choisis en fonction des objectifs pratiques du modèle de prédiction. Dans le cas des nomogrammes, on cherche à ce que le modèle distingue les patients les plus à risque de vivre l'événement de ceux moins à risque. Le score doit mesurer le pouvoir discriminant (*discrimination*) du modèle, qui est sa capacité à distinguer les patients à haut et bas risque. L'index de concordance (C-index) de Harrell [83] permet de mesurer ce pouvoir discriminant, et est le score préconisé pour la construction de nomogrammes [207, 12].

Le C-index évalue la capacité du modèle à prédire correctement l'ordre des événements de survie en tenant compte des données censurées. L'intuition derrière le C-index est la suivante : pour chaque paire de patients (i, j) , le modèle attribue un score de risque η_i et η_j . Si le modèle est performant, les patients avec des scores de risque plus élevés devraient avoir des temps de survie plus courts [205].

Pour calculer le C-index, on considère toutes les paires de patients (i, j) et on examine leurs scores de risque et leurs temps de survenue de l'événement :

- Si les deux temps T_i et T_j ne sont pas censurés, on peut observer quand les deux patients ont eu l'événement. La paire (i, j) est dite concordante si $\eta_i > \eta_j$ et $T_i < T_j$, et discordante si $\eta_i > \eta_j$ et $T_i > T_j$.
- Si les deux temps T_i et T_j sont censurés, on ne sait pas qui a vécu l'événement en premier, donc cette paire n'est pas prise en compte dans le calcul.
- Si l'un des temps T_i ou T_j est censuré, on observe seulement un événement. Supposons que T_i est observé et T_j est censuré :
 - Si $T_j < T_i$, on ne sait pas qui a vécu l'événement en premier, donc cette paire n'est pas prise en compte.
 - Si $T_j > T_i$, on sait que le patient i a vécu l'événement en premier. La paire (i, j) est alors concordante si $\eta_i > \eta_j$, et discordante si $\eta_i < \eta_j$.

Le C-index correspond au nombre de paires concordantes sur la somme des paires concordantes et discordantes :

$$\text{C-index} = \frac{\text{nombre de paires concordantes}}{\text{nombre paires concordantes} + \text{nombre de paires discordantes}}. \quad (3.19)$$

Ce qui se traduit par :

$$\text{C-index} = \frac{\sum_{i \neq j} 1\{\eta_i < \eta_j\} 1\{T_i > T_j\} d_j}{\sum_{i \neq j} 1\{T_i > T_j\} d_j}, \quad (3.20)$$

où $1\{\cdot\}$ est une fonction indicatrice qui vaut 1 si la condition est vraie, et 0 sinon.

Des valeurs de C-index proches de 0.5 indiquent que les prédictions de risque ne sont pas meilleures qu'un tirage aléatoire pour discriminer les patients à risque. Des valeurs proches de 1 indiquent un haut pouvoir discriminant.

Le C-index est une généralisation de l'aire sous la courbe ROC (AUC-ROC) qui peut prendre en compte les données censurées. Il représente une évaluation globale du pouvoir discriminant du modèle : c'est-à-dire la capacité du modèle à fournir un classement fiable des temps de survie en fonction des scores de risque individuels [205].

Dans les modèles de classification binaire, l'AUC-ROC est une métrique couramment utilisée pour évaluer les performances du modèle. L'AUC-ROC mesure la capacité du modèle à distinguer correctement entre les classes positives et négatives en traçant une courbe qui représente la sensibilité contre 1 moins la spécificité pour différents seuils de classification. L'AUC-ROC varie de 0.5 (aucune discrimination) à 1 (discrimination parfaite).

Dans l'étude de VAIDYA et al. [207], le C-index de QuRNOM est de 74%. Ainsi QuRNOM peut discerner dans 74% des cas, un patient qui connaîtra une rechute de son cancer dans les trois ans d'un patient sans rechute dans les trois ans. Ce nomogramme aide les oncologues à conseiller une chimiothérapie adjuvante aux patients. Si le nomogramme prédit une rechute du cancer, il oriente plutôt la décision vers la prescription de chimiothérapie, et inversement. Cela permet d'éviter un traitement lourd pour les patients qui sont à faible risque de rechuter.

Dans l'étude de JIANG et al. [106], le C-index du score HALP pour prédire la SG est de 78% et de 72% pour prédire la SSP. Le pouvoir discriminant du score HALP est élevé. Cependant ces résultats sont à prendre avec des pincettes car l'étape entraînement-validation n'est pas satisfaite.

3.3 Bilan : quels outils pour extraire et prédire les réponses aux chimiothérapies ?

Ce chapitre a présenté des techniques d'extraction de connaissances et de prédiction d'événements à partir des données hospitalières. Nous nous sommes concentrés sur celles que nous avons utilisées pour extraire les connaissances des réponses aux chimiothérapies.

L'extraction de connaissances permet de transformer des données brutes en informations exploitables de façon automatique. (*cf. section 3.1*)

Nous avons vu que les techniques de Traitement Automatique des Langues (TAL), comme la reconnaissance d'entités nommées (REN) et l'extraction de relations, jouent un rôle clé dans la structuration et l'analyse des données textuelles non structurées. (*cf. section 3.1.1*) Puis nous avons mis en évidence le rôle de mesures dans l'extraction des connaissances. Les mesures sont omniprésentes dans les disciplines scientifiques pour détecter des signaux et extraire des connaissances. Dans le TAL les distances d'édition permettent la reconnaissance d'entités nommées, en bio-informatique, les mesures de similarité permettent d'aligner des séquences biologiques et dans la recherche médicale, des indicateurs sont constamment utilisés pour évaluer des résultats cliniques. (*cf. section 3.1.2*)

Dans une seconde section, nous avons montré comment les modèles de survie sont utilisés pour prédire des événements d'intérêt clinique. (*cf. section 3.2*)

Nous avons d'abord présenté l'analyse de survie classique et son utilisation dans les études cliniques pour quantifier l'impact de facteurs de risques sur la survenue d'événements. Puis,

nous nous sommes intéressés à l'utilisation des régression logistique pour modéliser la survenue d'événement. Nous avons montré que le modèle de temps discret était à privilégier pour modéliser la survenue d'un événement avec une régression logistique. Nous avons comparé les différents modèles de survie exposés, et conclu que le modèle de Cox semblait le plus pertinent. (*cf.* sections 3.2.1, 3.2.2, 3.2.3)

Nous avons ensuite montré comment utiliser ces modèles pour prédire des événements. Pour illustrer le développement de modèles de prédiction, nous avons montré comment construire un nomogramme, outil d'aide à la décision clinique très répandu en oncologie. Les grandes étapes incluent la définition d'une question clinique claire, la sélection de variables explicatives pour définir un modèle de prédiction, l'entraînement du modèle et l'interprétation de ses performances. (*cf.* section 3.2)

Troisième partie

Contributions

CHAPITRE 4

REPRÉSENTER LES CHIMIOTHÉRAPIES ET LEURS RÉPONSES AVEC CHEMOONTO ET ONTOToX

4.1	Représenter les chimiothérapies et leurs réponses avec des ontologies	97
4.1.1	Pour standardiser les concepts relatifs aux chimiothérapies et leurs réponses	97
4.1.2	Pour les connecter à des modèles de connaissances de référence	100
4.1.3	Discussion sur les choix de représentation	102
4.2	Extraire les connaissances sur les chimiothérapies et leurs réponses depuis les entrepôts de données	106
4.2.1	Extraire les connaissances sur les chimiothérapies	106
4.2.2	Extraire les connaissances sur la survenue de toxicités	114
4.2.3	Discussion sur les méthodes d'extraction d'informations	116
4.3	Utiliser les ontologies instanciées comme des bases de connaissances	119
4.3.1	Instanciation de ChemoOnto	119
4.3.2	Instanciation d'OntoTox	120
4.4	Utiliser les ontologies pour découvrir de nouvelles connaissances	122
4.4.1	Raisonnement temporel	122
4.4.2	Comparer les schémas thérapeutiques avec ChemoKG et les plongements de graphes	125
4.5	Bilan	126



Le chapitre 1 a soulevé deux problèmes concernant la représentation et l'intégration des connaissances sur les chimiothérapies et leurs réponses :

- Les chimiothérapies et leurs réponses sont des domaines qui font intervenir des notions complexes à représenter et un vocabulaire spécifique (cycles, lignes, grade de toxicités,...) (sections 1.1 et 1.2 du chapitre 1.4.7)
- Les connaissances sur ces chimiothérapies et leurs réponses sont dispersées dans diverses sources de données et ne sont pas directement accessibles (sections 1.2.1 du chapitre 1)

Dans le même temps, le chapitre 2 a montré que les ontologies permettent de faciliter la structuration de connaissances complexes, leur intégration et leur analyse.

Dans ce chapitre, nous présentons la construction de deux ontologies pour représenter les chimiothérapies et leurs réponses :

- Le développement de ChemoOnto, une ontologie pour représenter le déroulement des chimiothérapies
- Le développement d'OntoTox, une ontologie pour représenter les toxicités survenues lors des chimiothérapies, extraites à partir de diverses sources.

Nous montrons qu'elles facilitent l'intégration et l'analyse des connaissances sur les chimiothérapies et leurs réponses extraites à partir des entrepôts de données.

Dans les deux cas, nous montrons la pertinence d'utiliser une ontologie en utilisant les classes et propriétés pour représenter des domaines complexes. (Section 4.1.1)

Pour les deux ontologies, nous montrons l'intérêt de réutiliser des modèles de données existants. Dans ChemoOnto, les liens avec la Time Ontology et Romedi permettent de représenter de façon précise la temporalité et de normaliser le nom des médicaments. Dans OntoTox, les liens avec le méthatésaurus de l'UMLS, la terminologie MedDRA et l'ontologie PROV-O permettent d'unifier les toxicités provenant de diverses sources.(Section 4.1.2)

Dans une deuxième section, nous présentons les méthodes utilisées pour extraire les connaissances sur les chimiothérapies et les survenues de toxicités. (Section 4.2)

Dans la troisième section, nous montrons que les instanciations de ChemoOnto et OntoTox avec les extractions constituent des bases de connaissances d'intérêt pour étudier les chimiothérapies et leurs réponses. (Section 4.3)

Dans la quatrième section nous montrons deux exemples d'utilisation de ChemoOnto et OntoTox comme des systèmes à base de connaissances avec le raisonnement temporel et l'utilisation d'algorithmes de plongement de graphes pour faire un clustering des schémas thérapeutiques. (Section 4.4)

Les résultats présentés dans ce chapitre ont fait l'objet d'une présentation orale à la conférence ISMB/ECCB Lyon 2023 [170], et de publication dans des conférences internationales avec comité de lecture : MedInfo 2021 [169] et SWAT4HCLS [105].

Préambule

Nous avons choisi un plan commun pour présenter simultanément deux ontologies, ChemoOnto et OntoTox, qui représentent des domaines différents bien que liés. De cette manière, nous tentons de convaincre le lecteur de la pertinence de représenter les connaissances médicales à l'aide de graphes, en ayant un plan guidé par les opportunités des représentations ontologiques, plutôt que de scinder le contenu en deux parties distinctes selon les domaines représentés. Néanmoins, ces deux domaines soulèvent des défis d'extraction, d'intégration et de représentation différents. Cette distinction est directement liée aux sources de données à partir desquelles les connaissances sont extraites (*cf.* sections 1.4.6.1.1, 1.4.6.1.2 et 1.4.7 du chapitre 1).

Dans le cas d'OntoTox, ontologie dédiée aux toxicités des chimiothérapies, les sources

de connaissances sont structurées (questionnaires), semi-structurées (tableaux des comptes rendus) et non structurées (texte libre des comptes rendus). En revanche, les sources de connaissances de ChemoOnto, ontologie dédiée au déroulement des chimiothérapies, sont entièrement structurées (enregistrements du logiciel Chimio). De ce fait, les défis à relever avec OntoTox se concentrent davantage sur les méthodes d'extraction, tandis qu'ils se concentrent plus sur les choix de représentation d'un processus complexe dans ChemoOnto. Ainsi, par rapport à OntoTox, ChemoOnto occupe une plus grande part de la discussion sur les choix de représentation (*cf. section 4.1.3*), tandis qu'OntoTox occupe une plus grande part de la discussion sur les méthodes d'extraction (*cf. section 4.2.3*).

4.1 Représenter les chimiothérapies et leurs réponses avec des ontologies

4.1.1 Pour standardiser les concepts relatifs aux chimiothérapies et leurs réponses

4.1.1.1 Définitions de classes et de propriétés pour qualifier les chimiothérapies dans ChemoOnto

Les traitements de chimiothérapie sont complexes et impliquent de nombreuses variables, comme nous avons pu le constater dans le chapitre 2 (cycles, lignes, médicaments anticancéreux, schémas thérapeutiques,...). Nous avons défini un ensemble de classes, de propriétés d'objets et de propriétés de données nécessaires pour représenter de manière complète le cours d'une chimiothérapie. ChemoOnto, se compose ainsi de six classes, dix propriétés d'objets et trente et une propriétés de données qui sont illustrées dans la figure 4.1 et la table en deux parties 4.1 et 4.2).

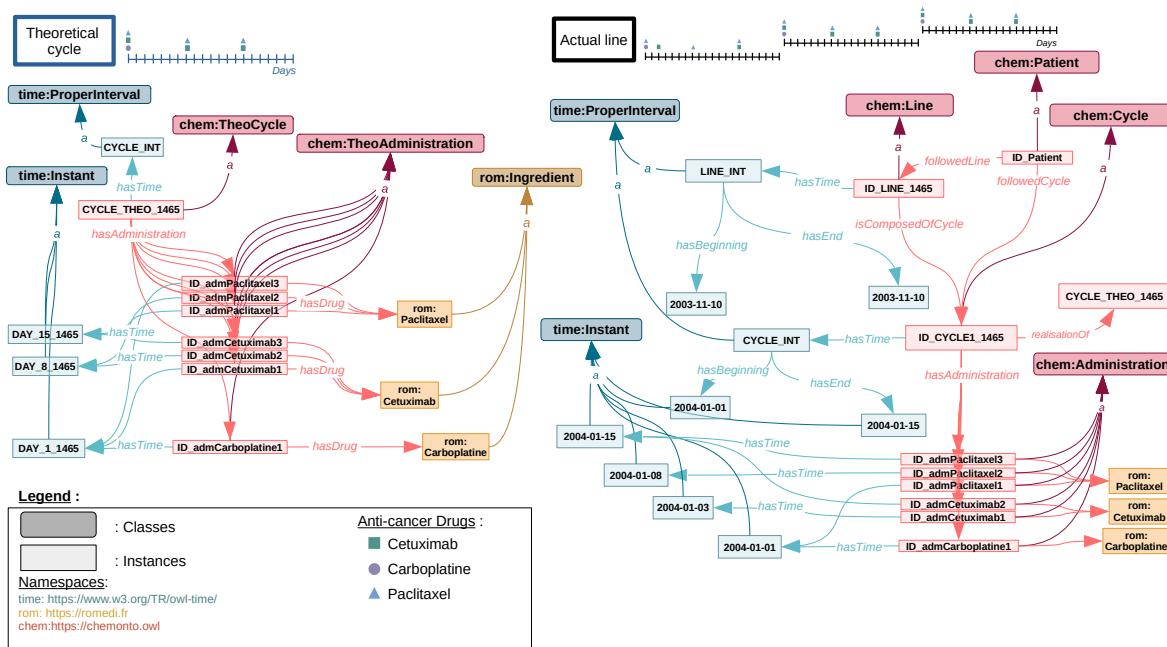


FIGURE 4.1 – Structure de ChemoOnto, illustrée avec un exemple de son instanciation avec un cycle standard (gauche) et une chimiothérapie suivie par un patient (droite)

Le graphe créé à partir de ChemoOnto est divisé en deux parties distinctes. La première partie représente les cycles théoriques, c'est-à-dire les cycles définis selon les schémas thérapeutiques. Deux classes et une propriété d'objet sont définis pour les représenter : les classes "TheoCycle" et "TheoDrugAdministration", et la propriété d'objet "hasTheoDrugAdministration" qui lie les instances de ces deux classes. Chaque instance de la classe "TheoCycle" représente précisément le cycle défini par un schéma thérapeutique en étant lié, via des instances de la propriété d'objet "hasTheDrugAdministration" à une combinaison d'instances de la classe "TheoAdministration". Les instances de la classe "TheoAdministration" complètent la description précise du cycle défini par le schéma thérapeutique en instanciant des propriétés d'objet vers des ontologies externes (pour le temps et les médicaments) ou de données pour décrire spécifiquement les modalités d'administration.

La seconde partie, quant à elle, permet de représenter le traitement réellement reçu par le patient. Nous avons définis quatre classes et sept propriétés d'objet (dont trois couples propriété \Rightarrow propriété inverse) pour représenter les traitements réellement suivis : les instances de la classe "Patient" sont liées aux instances de classes "Cycle" et "Line" via les instances de propriétés d'objets "followedCycle" et "followedLine". Les instances des classes "Line" et "Cycle" sont elles mêmes liées via des instances de la propriété d'objet "isComposedOfCycle". De manière analogue aux instances des classes "TheoCycle" et "TheoDrugAdministration", les instances de la classe "Cycle" décrivant les cycles réalisés, sont liées à des instances de la classe "DrugAdministration", via l'instanciation de la propriété d'objet "hasDrugAdministration". Toutefois, les classes réelles se distinguent de leurs analogues théoriques par l'ensemble de propriétés d'objet et de données dont leurs instances peuvent être sujets. Les propriétés de données des instances de la classe "DrugAdministration" ont notamment pour objectif de rester fidèles aux données extraites (cf. section 4.2).

Les deux parties sont reliées grâce à la propriété d'objet "realisationOf" (et sa propriété inverse \Leftarrow "isRealisedIn"), ce qui rend aisée la comparaison d'un cycle théorique et du traitement réel associé.

4.1.1.2 Définitions de classes et de propriétés qualifier les toxicités et leurs sévérités dans OntoTox

OntoTox est composée de onze classes organisées autour de la classe centrale ChemoTherapyToxicity. OntoTox comprend également huit propriétés d'objet et de données qui servent à qualifier les toxicités. La classe ChemoTherapyToxicity peut être liée aux classes Grade, StartDate et Patient grâce à des propriétés d'objet. La classe Grade possède sept sous-classes qui sont Grade0, Grade1, Grade2, Grade3, Grade4, Grade5 et GradeNull. GradeNull correspond à l'absence de grade détecté, tandis que Grade0 indique explicitement

Sujet : instances des classes listées	Prédicats : instances des propriétés listées	Objets : instances des classes listées ou littéraux	Description
Cycle	hasCycleNum	<i>entier</i>	Numéro du cycle dans la ligne
	hasCycleRegimenkey	<i>entier</i>	Clé schéma thérapeutique correspondant
	hasAdministration	DrugAdministration	Lien avec une administration réalisée
	realisationOf (\Leftarrow isRealisedIn)	TheoCycle	Lien avec le cycle théorique
	hasTime	ProperInterval	Lien avec l'intervalle de temps du cycle réalisé
TheoCycle	hasCycleRegimenkey	<i>entier</i>	Clé du schéma thérapeutique
	hasTheoAdministration	TheoDrugAdministration	Lien avec une administration théorique
	hasTime	ProperInterval	Lien avec l'intervalle de temps du cycle théorique
DrugAdministration	hasDose	<i>flottant</i>	Dose de médicament administrée (en mg)
	hasUnitOfDosage	<i>chaîne de caractères</i>	Unité du dosage administrée (ex : mg/m ² , mg/kg)
	isBolus	<i>booléen</i>	Indique un mode d'infusion en bolus (true/false)
	hasDoseNorm	<i>flottant</i>	Dose normalisée du médicament administré (ex : mg/m ² , mg/kg)
	hasCancerLocation	<i>chaîne de caractères</i>	Localisation du cancer traité
	hasReductionReason	<i>chaîne de caractères</i>	Motif renseigné si réduction de dose
	hasReductionPerc	<i>flottant</i>	Pourcentage de réduction renseigné
	hasPatBS	<i>flottant</i>	Surface corporelle du patient (en m ²)
	hasPatWeight	<i>flottant</i>	Poids du patient (en kg)
	hasCreat	<i>flottant</i>	Créatinine du patient (en $\mu\text{mol L}^{-1}$)
	hasMissingValues	<i>chaîne de caractères</i>	Valeurs manquantes pour calculer une dose normalisée
	hasInfusionDurationInMinutes	<i>entier</i>	Durée de l'infusion en minutes
	hasRoute	<i>chaîne de caractères</i>	Voie d'administration du médicament (ex : IV, SC)
	hasDrugRole	<i>chaîne de caractères</i>	Rôle du médicament (ex : anti cancer, anti effet indésirable)
	hasDrugNameInDB	<i>chaîne de caractères</i>	Nom du médicament dans la base de données
	hasDrugKeyInDB	<i>entier</i>	Clé du médicament dans la base de données
	hasRomediIngredient	Ingrédients de Romedi	Lien avec l'ingrédient associé au médicament administré
	hasTime	Instant	Lien avec la date d'administration

TABLE 4.1 – Triplets possibles dans ChemoOnto. Les entités de ChemoOnto sont surlignées en rouge, celles de la Time Ontology en bleu et celles de Romedi en orange. Les littéraux sont écrits en italique.

CHAPITRE 4

Sujet :	Prédicats :	Objets :	Description
instances des classes listées	instances des propriétés listées	instances des classes listées ou littéraux	
TheoDrugAdministration	hasDosage	<i>flottant</i>	Dosage théorique du médicament
	hasUnitOfDosage	<i>chaîne de caractères</i>	Unité du dosage théorique (ex : mg/m ² , mg/kg)
	hasDayDrugAdmSinceCycleStart	<i>entier</i>	Jour théorique d'administration du médicament dans le cycle
	isBolus	<i>bool</i>	Indique un mode d'infusion théorique en bolus (true/false)
	hasInfusionDurationInMinutes	<i>entier</i>	Durée théorique de l'infusion en minutes
	hasRoute	<i>chaîne de caractères</i>	Voie d'administration du médicament (ex : IV, SC)
	hasDrugRole	<i>chaîne de caractères</i>	Rôle du médicament (ex : anti cancer, anti effet indésirable)
	hasDrugNameInDB	<i>chaîne de caractères</i>	Nom théorique du médicament dans la base de données
	hasDrugKeyInDB	<i>entier</i>	Clé théorique du médicament dans la base de données
	hasRomediIngredient	Ingrediénts de Romedi	Lien avec l'ingrédient associé au médicament de l'administration théorique
	hasTime	Instant	Lien avec un jour théorique
Line	hasRegimenName	<i>chaîne de caractères</i>	Nom du schéma thérapeutique
	hasLineRegimenkey	<i>entier</i>	Clé du schéma thérapeutique
	isComposedOfCycle (= cycleOfLine)	Cycle	Cycles composant la ligne de traitement
	hasTime	ProperInterval	Lien avec l'intervalle de temps associé à la ligne
Patient	hasBirthDate	<i>chaîne de caractères</i>	Date de naissance du patient
	hasSexe	<i>chaîne de caractères</i>	Sexe du patient
	followedCycle (= cycleIsFollowedBy)	Cycle	Cycles suivis par le patient
	followedLine (= lineIsFollowedBy)	Line	Ligne de traitement suivie par le patient
Intervales de cycles réalisés	hasBeginning	Instant	Lien entre l'instant de début et l'intervalle de cycle réalisé
	hasEnd	Instant	Lien entre l'instant de fin et l'intervalle de cycle réalisé
Instants de début de cycles réalisés	inXSDDate	<i>xsd :date</i>	Date de début de cycle au format ISO 8601
Instants de fin de cycles réalisés	inXSDDate	<i>xsd :date</i>	Date de fin de cycle au format ISO 8601
Intervales de cycles théoriques	hasXSDDuration	<i>xsd :duration</i>	Durée du cycle en jours au format ISO 8601
Instants d'administrations réalisées	inXSDDate	<i>xsd :date</i>	Date d'administration au format ISO 8601
Instants d'administrations théoriques	inside	intervalles de cycles théoriques	Lien entre jours d'administration et intervalles cycles théoriques
Intervales de lignes	hasBeginning	Instant	Lien entre l'instant de début et l'intervalle de ligne
	hasEnd	Instant	Lien entre l'instant de fin et l'intervalle de ligne
Instants de début de ligne	inXSDDate	<i>xsd :date</i>	Date de début de ligne au format ISO 8601
Instants de fin de ligne	inXSDDate	<i>xsd :date</i>	Date de fin de ligne au format ISO 8601

TABLE 4.2 – Suite des triplets possibles dans ChemoOnto. Les entités de ChemoOnto sont surlignées en rouge, celles de la Time Ontology en bleu et celles de Romedi en orange. Les littéraux sont écrits en italique.

un grade 0, c'est-à-dire l'absence de cette toxicité, ce qui est couramment observé dans les questionnaires et les tableaux des données patients. Les instances de ChemotherapyToxicity peuvent être associées à différentes propriétés de données pour caractériser le contexte de l'extraction de la toxicité (par exemple, isNeg, isHyp qualifient le fait que la toxicité peut être extraite en tant que fait négatif ou hypothétique).

4.1.2 Pour les connecter à des modèles de connaissances de référence

Un des objectifs d'utiliser des ontologies pour représenter les chimiothérapies et leurs réponses est également d'utiliser des ontologies existantes et standards. Cette section présente les différentes réutilisations d'ontologies de référence.

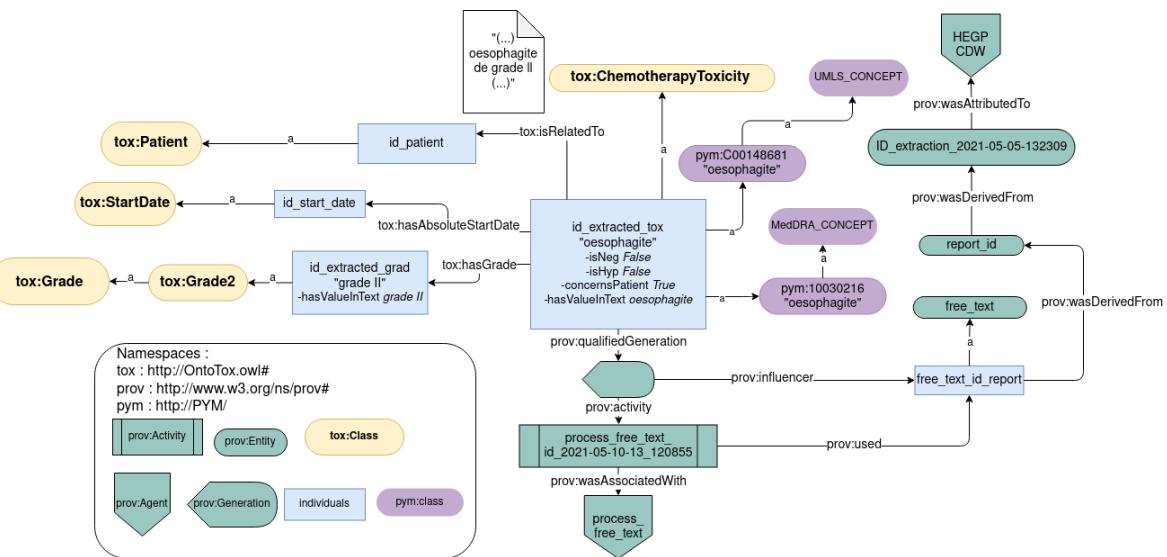


FIGURE 4.2 – Structure d’OntoTox illustrée avec un exemple de son instantiation depuis une extraction du texte libre : “œsophagite de grade II”

4.1.2.1 Utilisation de modèles de référence pour représenter la temporalité et les médicaments dans ChemoOnto

Le temps et les médicaments représentés avec des modèles de données externes
La représentation précise d’un traitement de chimiothérapie nécessite la prise en compte de deux éléments fondamentaux : le temps et les médicaments. Des modèles de données existants couvrent ces deux domaines. Il est pour cette raison intéressant de réutiliser ces représentations.

La Time Ontology pour représenter le temps Pour représenter le temps, nous avons utilisé la Time Ontology, qui est une ontologie de référence pour les données temporelles (*cf.* section 2.3.4.1) [45]. En réutilisant cette ontologie dans ChemoOnto, nous avons représenté de manière précise les intervalles de temps correspondants aux cycles et aux lignes suivies et aux cycles théoriques, à l’aide de la classe **time:ProperInterval**. Les intervalles de cycles et lignes suivis sont délimités par des dates de début et de fin, tandis que les intervalles de temps des cycles théoriques sont définis de manière relative, avec une durée spécifique au schéma thérapeutique. Chaque administration réelle est associée à une date instanciant la classe **time:Instant** tandis que les jours d’administration théoriques sont là encore définis de manière relative.

On remarque que même si nous réutilisons la Time Ontology pour représenter la temporalité dans ChemoOnto, nous ne réutilisons pas totalement le modèle *4D-fluent* décrit par BATSAKIS et al. [15] (*cf.* section 2.3.4.1). Dans la figure 2.1 qui présente un exemple du modèle *4D-fluent*, on constate que les intervalles de temps du contrat de l’employé (“EmployeeTimeSlice”) et de vie de l’entreprise (“CompanyTimeSlice”) sont directement les sujets des prédictats “worksFor” et “employs”. À titre d’exemple, si on fait l’analogie avec les lignes et cycles de ChemoOnto, pour être en adéquation avec le modèle *4D-fluent*, la propriété

"isComposedOfCycles" devrait lier des instances entre les instances d'intervalles de lignes et de cycles plutôt qu'entre les instances de lignes et les cycles. Le modèle *4D-fluent* a pour objectif de suivre les évolutions des instances de manière dynamique. ChemoOnto a pour objectif de retracer l'histoire de chimiothérapie des patients. Dans le cas de ChemoOnto, nous pensons qu'il est plus pertinent que les entités temporelles instantient exclusivement des propriétés temporelles. On est cependant plus proche du modèle *4D-fluent* que du modèle de réification (*cf.* figure 2.1), puisque chacune des instances de ligne et de cycle génère une entité temporelle.

Romedi pour représenter les médicaments Nous avons lié les médicaments de ChemoOnto aux ingrédients de la base de connaissances Romedi [42], spécifiquement conçue pour la détection des médicaments en français. Ce lien nous a permis de représenter avec précision les médicaments administrés lors du traitement de chimiothérapie. De plus, chaque ingrédient de Romedi est lié au système de classification internationale des médicaments ATC (Anatomical Therapeutic Chemical), qui offre un système normalisé d'identification et de classification des médicaments (*cf.* section 2.3.4.3 du chapitre 4).

4.1.2.2 Utilisation de modèles de référence pour unifier les connaissances sur les toxicités et qualifier leurs provenances dans OntoTox

L'un des objectifs d'OntoTox est d'unifier les connaissances sur les toxicités provenant de différentes sources de données. Pour atteindre cet objectif, OntoTox valide deux pré-requis. Le premier pré-requis est de guider la normalisation des toxicités extraites. Le deuxième pré-requis concerne l'encodage des informations de provenance.

Les classes de l'UMLS et de MedDRA pour normaliser les toxicités Il peut y avoir des variations dans les termes utilisés pour décrire les toxicités et leur gravité. La normalisation des toxicités avec des concepts de l'*Unified Medical Language System* (UMLS) et de la *Medical Terminology for Regularory Activities* (MedDRA) permet d'harmoniser ces variations. L'UMLS et MedDRA sont des terminologies de références dans le domaine médical qui utilisent des identifiants pour décrire des concepts médicaux. (*cf.* sections 1.2.3 du chapitre 1 et 2.3.4.3 du chapitre 2)

PROV-O pour encoder la provenance Comme les toxicités sont extraites à partir de sources distinctes, il est essentiel de connaître l'origine de ces données pour assurer leur traçabilité. OntoTox utilise l'ontologie PROV-O pour encoder les informations de provenance . Cela facilite la transparence et la reproductibilité des analyses effectuées à partir de ces données. (*cf.* section 2.3.4.2 du chapitre 2)

4.1.3 Discussion sur les choix de représentation

Dans cette section, nous allons discuter des choix de représentation, principalement ceux effectués dans ChemoOnto, en mettant en lumière les raisons pour lesquelles certaines classes et propriétés d'ontologies ou de terminologies existantes n'ont pas été réutilisées. Nous présenterons également un tableau récapitulatif pour illustrer les alternatives possibles et les justifications de nos choix.

4.1.3.1 Tableau récapitulatif des alternatives de représentation avec des ontologies existantes

Ce tableau résume les classes et propriétés de données de ChemoOnto et OntoTox et les alternatives possibles dans les modèles de données existants.

Entité de ChemoOnto ou d'OntoTox	Ontologie alternative	Entité dans l'ontologie alternative
TheoCycle	NCIT	NCIT :C15697 - Treatment Regimen, NCIT :C60735 - Treatment Plan
	OBI	OBI :0003069 - Sequencing protocol
chem :Cycle	NCIT	NCIT :C15697 - Treatment Regimen
chem :DrugAdministration	OBI	OBI :0000299 - Administration
	SNOMED_CT	SNOMED :73639000 - Administration
	DRON	DRON :00000031 - Drug Administration
	PDRO	PDRO :0010008 - prescribed drug administration
chem :hasAdministration	OBI	OBI :0000299 - has_specified_output
chem :hasTheoAdministration	OBI	OBI :0000299 - has_specified_output
chem :isComposedOfCycle	RDF Schema	rdf :Seq
chem :hasUnitSTR	UO	UO :0000022 - milligram
chem :hasDosage	SNOMED_CT	SNOMED :732943007 - Dosage
	OBI	OBI :0001919 - Dosage
chem :realisationOf	OBI	OBI :0001874 - Protocol Execution
chem :hasDose	RxNorm	RxNorm : dose
chem :isBolus	SNOMED_CT	SNOMED :70694002 - Bolus dose
chem :hasCancerLocation	NCIT	NCIT :C12264 - Cancer Location
chem :hasInfusionDurationInMinutes	LOINC	LOINC :33879-9 - Duration of Infusion
	TO	time :hasXSDDuration
chem :hasRoute	SNOMED_CT	SNOMED :263491000 - Route of administration
chem :hasDrugRole	NCIT	NCIT :C37938 - Drug Role
tox :ChemotherapyToxicity	NCIT	NCIT :C42606 - Chemotherapeutic Agent Toxicity
tox :Grade	NCIT	NCIT :C42606 - Toxicity Grade

TABLE 4.3 – Comparaison des classes et propriétés de ChemoOnto avec des ontologies existantes

4.1.3.2 Compatibilité avec d'autres objectifs de recherche

ChemoOnto et OntoTox ont été développés en parallèle d'autres projets de recherche, en particulier celui de ProtoDrift (cf. chapitre 5). Ces objectifs de recherche ont influencé des choix de représentation pragmatiques pour assurer une intégration et une utilisation efficaces. Des graphes fonctionnels et adaptables ont été priorisés sur une conformité stricte aux standards existants, et parfois sur une représentation plus fidèle à la réalité.

Pour illustrer cette problématique de compatibilité avec le développement parallèle d'autres projets, nous allons expliquer nos choix de représentations temporelles des administrations d'anticancéreux, et de survenue de toxicités.

Parmi les alternatives de représentation, le tableau 4.3 propose des propriétés pour représenter le temps d'infusion des administrations avec la Time Ontology ou l'ontologie LOINC.

Dans ChemoOnto, nous avons lié les instances d'administration à des instances de la classe Instant de la Time Ontology, tant pour les instances d'administrations théoriques (TheoDrugAdministration) que suivies (DrugAdministration). Or une représentation plus fidèle à la réalité lierait les administrations à des durées d'infusion, et donc plutôt à des intervalles (instances de la classe time :ProperInterval) qu'à des instants (instances de la classe time :Instant). Il n'y aurait ainsi plus les propriétés de données "hasInfusionDurationInMinutes" pour les instances des classes TheoDrugAdministration et DrugAdministration, les premières ayant une représentation relative du temps d'infusion et les secondes absolue, avec des dates de début et de fin représentés au format xsd :dateTime plutôt qu'au format xsd :date. Nous avons restreint le niveau de granularité de la description temporelle au niveau du jour car nous n'avions pas besoin d'un niveau plus élevé pour calculer ProtoDrift.

Dans OntoTox, ce sont aussi des instances de la classe time :Instant qui sont liées aux extractions de toxicités. Les dates des instants associés aux extractions de toxicités sont des métadonnées, elles sont associées à l'édition de la source d'extraction (date d'édition du questionnaire ou du compte rendu), qui ne correspond pas forcément à la date de l'occurrence de la toxicité extraite. Or, il serait plus juste là aussi d'instancier des intervalles représentant la durée de la toxicité. Cependant cela implique de développer des méthodes d'extraction de ces dates dans les différentes sources de toxicités, ce qui est discuté dans la section 4.2.3.

4.1.3.3 Difficulté à intégrer diverses classes existantes

Bien qu'il existe de nombreuses classes pour représenter différentes parties des graphes, les réunir de manière cohérente peut être complexe et chronophage. ChemoOnto a été conçue pour être une solution intégrée qui simplifie ce processus. La multiplicité des terminologies et ontologies dans le domaine médical rend la standardisation difficile, et ces terminologies ne sont pas toujours compatibles entre elles. L'un des objectifs de représentation de ChemoOnto est la comparaison entre un cycle théorique défini exactement par un schéma thérapeutique et un cycle réalisé. Aucune combinaison de classes existantes dans les ontologies médicales ne permet à notre sens d'atteindre cet objectif.

4.1.3.4 Compromis entre réutilisation et expressivité

La réutilisation de classes existantes peut limiter l'expressivité et la spécificité des représentations [73]. ChemoOnto permet une plus grande flexibilité pour adapter les cycles de chimiothérapie théoriques et suivis aux besoins spécifiques du domaine. Par exemple, pour représenter les jours de traitement depuis le début du cycle théorique, la propriété *hasDay-DrugAdmSinceTheCycleStart* est utilisée car il n'existe pas de propriété appropriée dans la Time Ontology pour spécifier le nombre de jours relatifs depuis le début d'un intervalle.

4.2 Extraire les connaissances sur les chimiothérapies et leurs réponses depuis les entrepôts de données

4.2.1 Extraire les connaissances sur les chimiothérapies

En réalité, les schémas thérapeutiques et le déroulement des chimiothérapies peuvent être représentés sous forme tabulaire avec des requêtes SQL sur les entrepôts de données. (*cf. section 1.4.6*). Avec ChemoOnto, nous proposons une méthode alternative de représentation, qui nous semble être plus adaptée aux besoins de notre étude. Dans les deux sections suivantes, nous montrons comment l'on passe de représentations tabulaires extrayables des entrepôts de données, aux représentations ontologiques et tabulaires de ChemoOnto. Le projet gitlab [ChemoOntoTox](#) permet d'effectuer de telles transformations pour cinq faux patients.

4.2.1.1 Reconstruire les schémas thérapeutiques théoriques

Les connaissances sur les schémas thérapeutiques théoriques sont présentes dans quatre tables issues de la base de données du logiciel Chimio et intégrées dans l'entrepôt de données (*cf. section 1.4.6.1.1*). Une première requête permet de joindre les informations sur les schémas thérapeutiques de la table "PROTOCOLE" aux informations sur chacune des administrations des schémas présentes dans la table "LIGNEPROTOCOLE". La figure 4.3 schématisse la structure de la table de résultats de cette requête. Une deuxième requête permet d'obtenir la correspondance entre les clés de molécules anticancéreuses et leurs noms en sélectionnant les deux colonnes correspondantes depuis la table "DCI" (*cf. figure 4.4*).

REPRÉSENTER LES CHIMIOTHÉRAPIES ET LEURS RÉPONSES AVEC CHEMOONTO ET ONTOTox

FIGURE 4.3 – Structure des résultats de la requête pour obtenir le déroulement des schémas thérapeutiques théoriques. Les points "..." indiquent que les résultats continuent.

Chaque schéma thérapeutique possède une clé (colonne "clé du schéma thérapeutique"), un nom (colonne "Nom du schéma thérapeutique") et une durée de cycle en jours (colonne "Durée du cycle"). Chaque entrée de cette table concerne pour un schéma thérapeutique (colonnes "clé du schéma thérapeutique" / "Nom du schéma thérapeutique"), une molécule anticancéreuse ("Clé de la molécule") et un dosage (colonne "dosage") donnés, les autres modalités d'administrations telles que le code d'unité (colonne "Unité"), la voie d'administration (colonne "Voie d'administration"), la durée d'administration (colonnes 'Durée de l'administration en heures complètes" et "Durée de l'administration en minutes complémentaires") et le ou les jour(s) d'administrations dans le cycle (colonne "Jour(s) d'administration"). Comme évoqué précédemment, la correspondance entre clé de molécule et nom de molécule se fait avec deux colonnes de la table "DCI". Un schéma de la structure de la table de correspondance est présent ci-dessous (*cf. figure 4.4*).

Clé de la molécule	Nom de la molécule
30	Methotrexate
43	Vinblastine
49	Cisplatine
32	Paclitaxel
67	Pemetrexed
12	Doxorubicine
.	.
.	.

FIGURE 4.4 – Correspondance entre clé et nom de molécules. Les points "..." indiquent que les résultats continuent

Ces tables concentrent les informations nécessaires et suffisantes pour représenter les schémas thérapeutiques. Avec ces exemples, on peut représenter les schémas thérapeu-

tiques 1357 "TAXOL HEBDOMADAIRE J1, J8, J15, J22 J1=J28" et 1376 "M-VAC HD (Cisplatin)".

Nous apportons ci-dessous quelques précisions et remarques supplémentaires sur le format et l'extraction d'information à partir de ces tables.

Les code d'unité À notre connaissance, il n'existe pas dans la base de données du logiciel Chimio de table de correspondance entre les codes d'unité et les unités. Cette correspondance a donc été déterminée empiriquement en vérifiant, à l'aide du dictionnaire Vidal [110], les unités et les voies d'administration d'un ensemble de molécules, et en consultant les pharmaciens de l'HEGP pour confirmation.

Le schéma des jours d'administration Le schéma de jours d'administration est une séquence de caractères composée de "N" ou de "O". Chaque caractère dans la séquence correspond à un jour dans le cycle du schéma thérapeutique. La présence du caractère "O" à un indice donné indique que l'administration est prévue pour ce jour du cycle, tandis que la présence du caractère "N" indique qu'aucune administration n'est prévue ce jour-là. Par exemple, si la séquence est "ONNNNNNONNNNNNONNNNNNONNNNN", cela indique un cycle de 28 jours avec une administration prévue aux jours 1, 8, 15 et 22.

Le bolus Il faut noter qu'il est difficile de trouver une définition claire et précise de l'administration en bolus. Une administration en "bolus" fait référence à l'administration d'une dose élevée médicamenteuse sur une courte durée. Cette durée varie selon les sources, mais est généralement fixée à une durée inférieure à 30 minutes [130, 98]. Au vu des questionnements actuels au sein de la communauté des oncologues sur l'administration en bolus (*cf.* 1.2.6), il nous a semblé important de qualifier ce mode d'administration. Nous avons donc qualifier cette administration avec une règle sur la durée d'infusion (inférieure ou non à 30 minutes).

Les médicaments anti-effets indésirables Les schémas thérapeutiques des tables de la bases de données Chimio contiennent aussi les informations sur les administrations de médicaments anti-effets indésirables. Pour ces administrations, la valeur de la colonne "Code du médicament commercial" est alors différente de '-1' et c'est la table "PRODUIT" qui permet de faire le lien entre le code et le médicament. ChemoOnto offre aussi la possibilité de représenter ces administrations. Pour simplifier et parce que nous n'avons pas utilisé cette option dans la suite du travail, nous ignorons cette partie de la représentation.

Les particularités intra-hospitalières ChemoOnto a été déployée dans deux hôpitaux qui intègrent les données sur les schémas thérapeutiques de la base de données Chimio (*cf.* section 1.4.6). Bien que la structure de ces tables reste la même d'un hôpital à l'autre, les données qui y sont entrées, et particulièrement les clés des éléments, diffèrent d'un hôpital à l'autre. Les tables de correspondances clé-molécule anticancéreuse, clé-unité, clé-produit, clé-nom de schéma thérapeutique ne sont pas standardisées.

L'identification des médicaments avec Romedi Nous avons utilisé RomediApp [42] pour détecter les noms français des molécules attribuées dans la base de donnée, les standardiser et les lier au système international de classification thérapeutique de l'*Anatomical Therapeutic Classification System* (ATC) (*cf.* section 2.3.4.3.3).

On expose ci-dessous les représentations ontologique (*cf.* figure 4.5) et tabulaire (figure 4.6) de ChemoOnto du schéma thérapeutique 1357 présent dans la figure 4.3.

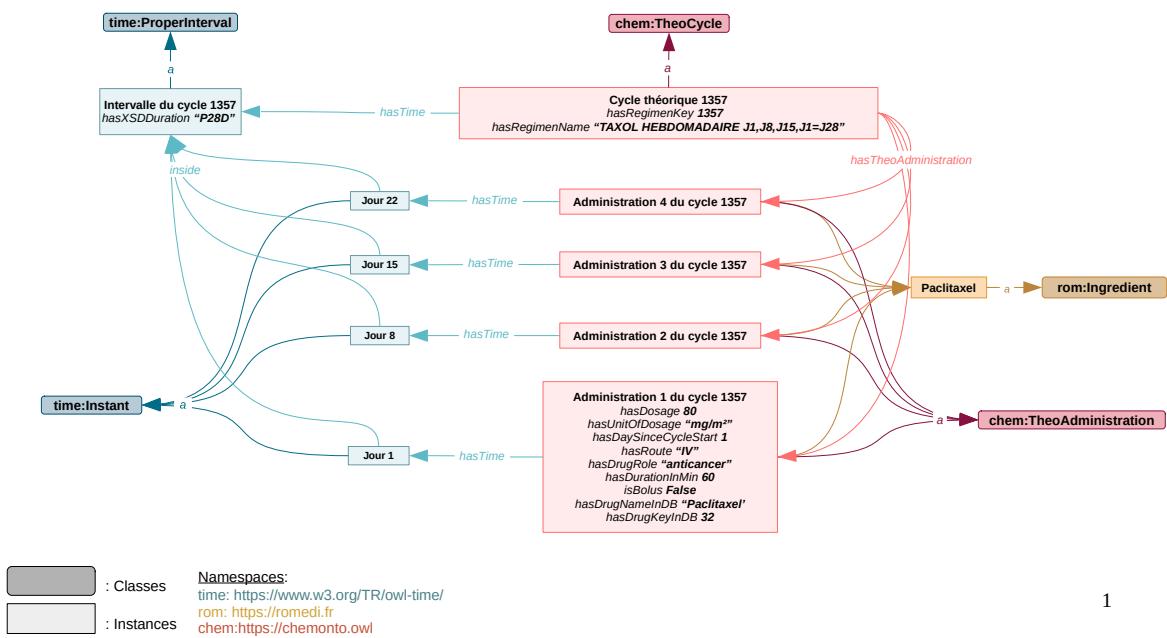


FIGURE 4.5 – Représentation ontologique du schéma thérapeutique 1357 avec ChemoOnto. Pour alléger la visualisation, les propriétés de données associées sont indiquées en italique directement dans les instances, plutôt qu'avec des flèches, et elles ne sont indiquées que dans une seule des quatre instances de "TheoDrugAdministration"

CHEMOMONOGRAPHIE

CAPITOLE 4

Clé de l'administration	Clé du schéma thérapeutique	Nom du schéma thérapeutique	Durée du cycle en jours	Ingrédient Romedi	Nom de l'ingrédient	Dosage	Unité	Voie d'administration	Durée de l'administration en minutes	Numéro du jour de l'administration dans le cycle	Mode d'infusion en bolus
ADM_0_1357_32	1357	TAXOL HEBDOMADAIRE J1, J8, J15, J22 J1=J28	28	INck2vkfdamu5ebj8i62fjq20pki8ltio	Paclitaxel	80	mg/m ²	IV	60	1	False
ADM_1_1357_32	1357	TAXOL HEBDOMADAIRE J1, J8, J15, J22 J1=J28	28	INck2vkfdamu5ebj8i62fjq20pki8ltio	Paclitaxel	80	mg/m ²	IV	60	8	False
ADM_2_1357_32	1357	TAXOL HEBDOMADAIRE J1, J8, J15, J22 J1=J28	28	INck2vkfdamu5ebj8i62fjq20pki8ltio	Paclitaxel	80	mg/m ²	IV	60	15	False
ADM_3_1357_32	1357	TAXOL HEBDOMADAIRE J1, J8, J15, J22 J1=J28	28	INck2vkfdamu5ebj8i62fjq20pki8ltio	Paclitaxel	80	mg/m ²	IV	60	22	False
ADM_0_1376_30	1376	M-VAC HD (Cisplatin)	14	IN1dr3uofk7desgg94qjcq093gejn0fkq	Methotrexate	30	mg/m ²	IV	30	1	False
ADM_0_1376_12	1376	M-VAC HD (Cisplatin)	14	INhvp8a3vq0m2vqomagqijif4mhgqu7p	Doxorubicine	30	mg/m ²	IV	15	1	True
ADM_0_1376_49	1376	M-VAC HD (Cisplatin)	14	INu1ijcoqh456snv1bvbefs3ns4ge6qq	Cisplatin	70	mg/m ²	IV	60	1	False
ADM_0_1376_43	1376	M-VAC HD (Cisplatin)	14	IN1c3pepvfr3s0feq4ed5r3fbgcchtehv	Vinblastine	3	mg/m ²	IV	15	1	True
.
.
.

FIGURE 4.6 – Représentation tabulaire du schéma thérapeutique 1357 avec ChemoOnto. Chaque ligne de la table correspond à une administration d'un schéma thérapeutique spécifique avec ses modalités.

4.2.1.2 Reconstruire le déroulement des chimiothérapies suivies

REPRÉSENTER LES CHIMIOTHÉRAPIES ET LEURS RÉPONSES AVEC CHEMOONTO ET ONTOTox

Identifiant du patient	Clé du schéma thérapeutique	Nom du schéma thérapeutique	Numéro du cycle	Date de l'administration	Localisation de la tumeur	Clé de la molécule administrée	Dose d'unité	Code	Date de naissance du patient	Sexe du patient	Surface corporelle	Taille	Poids	Créatinine	Durée d'administration	Voie d'administration	Motif de la réduction	Pourcentage de réduction
P1	1357	TAXOL HEBDOMADAIRE J1, J8, J15, J22 J1=J28	1	2004-11-18	VESSIE et UROTHELIAL	32	150.0	2	1955-09-17	F	1.74	158	66	116	1:00	Voie Intra Veineuse	None	None
			1	2004-11-25	VESSIE et UROTHELIAL	32	132.0	2			1.74	158	66	116	1:00	Voie Intra Veineuse	None	None
			1	2004-12-02	VESSIE et UROTHELIAL	32	132.0	2			1.74	158	66	116	1:00	Voie Intra Veineuse	None	None
			1	2004-12-09	VESSIE et UROTHELIAL	32	132.0	2			1.74	158	66	116	1:00	Voie Intra Veineuse	None	None
			2	2004-12-16	VESSIE et UROTHELIAL	32	132.0	2			1.74	158	66	116	1:00	Voie Intra Veineuse	None	None
			2	2004-12-23	VESSIE et UROTHELIAL	32	132.0	2			1.74	158	66	85	1:00	Voie Intra Veineuse	None	None
			2	2004-12-30	VESSIE et UROTHELIAL	32	132.0	2			1.74	158	66	85	1:00	Voie Intra Veineuse	None	None
			2	2005-01-06	VESSIE et UROTHELIAL	32	132.0	2			1.74	158	66	85	1:00	Voie Intra Veineuse	None	None
	1378	M-VAC HD (Cisplatin)	1	2006-08-01	VESSIE et UROTHELIAL	49	138	2			1.74	158	66	85	0:15	Voie Intra Veineuse	None	None
		
		
		
P2	1000	CISPLATINE 40mg/m ² J1J2 - ALIMTA 500mg/ m ² HDJ	1	2011-03-09	RESPIRATOIRE et THORACIQUE	67	900.0	2	1951-04-03	M	1.86	162	74	60	0:10	Voie Intra Veineuse	None	None
.	
.	
.	

FIGURE 4.7 – Structure des résultats de la requête pour obtenir le déroulement des lignes suivies de chimiothérapie. Les "|" indiquent une valeur identique à la dernière valeur observée de la colonne, tandis que les points "..." indiquent que les résultats continuent.

Les connaissances sur les administrations réelles de chimiothérapies sont essentiellement présentes au sein de la table "OBSERVATION_FACT" des entrepôts de données (*cf.* section 1.4.6.1.2). Pour reconstruire les lignes de chimiothérapies réellement suivies, il faut ordonner les administrations suivant leurs dates, par patient et par schéma thérapeutique suivi.

Le calcul de la dose normalisée Pour pouvoir comparer la dose administrée au dosage du schéma thérapeutique, il faut la normaliser suivant son unité de dosage. Le plus souvent l'unité est en mg/m² ou en mg/kg. C'est l'une des raisons pour lesquelles il est important de conserver le poids et la surface corporelle du patient au niveau de l'administration. Pour obtenir la dose normalisée, il faut donc souvent diviser la dose totale administrée par le poids ou la surface corporelle. Certaines molécules comme le carboplatine ont une unité de dosage spécifique. Le dosage du carboplatine se calcule à l'aide de la formule de Chatelut [33], et dépend de l'âge, du sexe et du taux de créatinine du patient.

Les valeurs manquantes Il arrive que certaines valeurs soient manquantes pour recalculer la dose normalisée (poids, surface corporelle, taille, créatinine). Si le dosage possède une unité en mg/m², que la surface corporelle est manquante mais pas le poids ni la taille, alors celle-ci est calculée avec la formule de Boyd[22]. Dans le cas où il reste des valeurs manquantes pour calculer la dose normalisée, nous avons attribué une valeur correspondant à, si elle existait, la valeur la plus récente non manquante du champ correspondant. Si le champ est toujours manquant pour un patient, alors la valeur de la dose normalisée est aussi manquante.

Les intervalles de temps associés aux cycles et aux lignes Le intervalles de cycles et de lignes sont construits à partir des dates d'administrations. Pour un patient donné, une date de début de ligne est attribuée à la date de la première administration enregistrée pour le schéma thérapeutique suivi. La date de fin de ligne correspond quant à elle à la date de dernière administration enregistrée pour le schéma thérapeutique. La date de début de cycle correspond à la première date enregistrée pour le numéro du cycle suivi. Excepté pour le dernier cycle de la ligne, la date de fin d'un cycle correspond à la veille de la date enregistrée pour la première administration du cycle suivant. Pour le dernier cycle de la ligne, la date de fin correspond, comme pour la ligne, à la date de la dernière administration enregistrée pour le schéma thérapeutique.

L'alignement entre nom et clé de schéma thérapeutique Dans l'entrepôt de données de l'HEGP, le nom des schémas thérapeutiques est présent dans la colonne "OBSERVATION_BLOB", mais la clé numérique correspondante est absente. Cette situation pose problème, car les tables qui contiennent les informations détaillées sur les schémas thérapeutiques ne présentent pas de relation bijective entre les noms des schémas thérapeutiques et leurs clés numériques. Autrement dit, certaines clés de schémas thérapeutiques sont associées à plusieurs noms de schémas, et donc à plusieurs schémas thérapeutiques distincts (relations 1-n). Lorsqu'un nom de schéma thérapeutique est associé à plusieurs clés possibles, nous avons attribué arbitrairement une clé parmi les clés possibles. Cette méthode permet de maintenir une certaine cohérence dans les données, bien qu'elle puisse introduire des biais potentiels. Un processus plus approfondi et systématique pourrait améliorer cette correspondance à l'avenir.

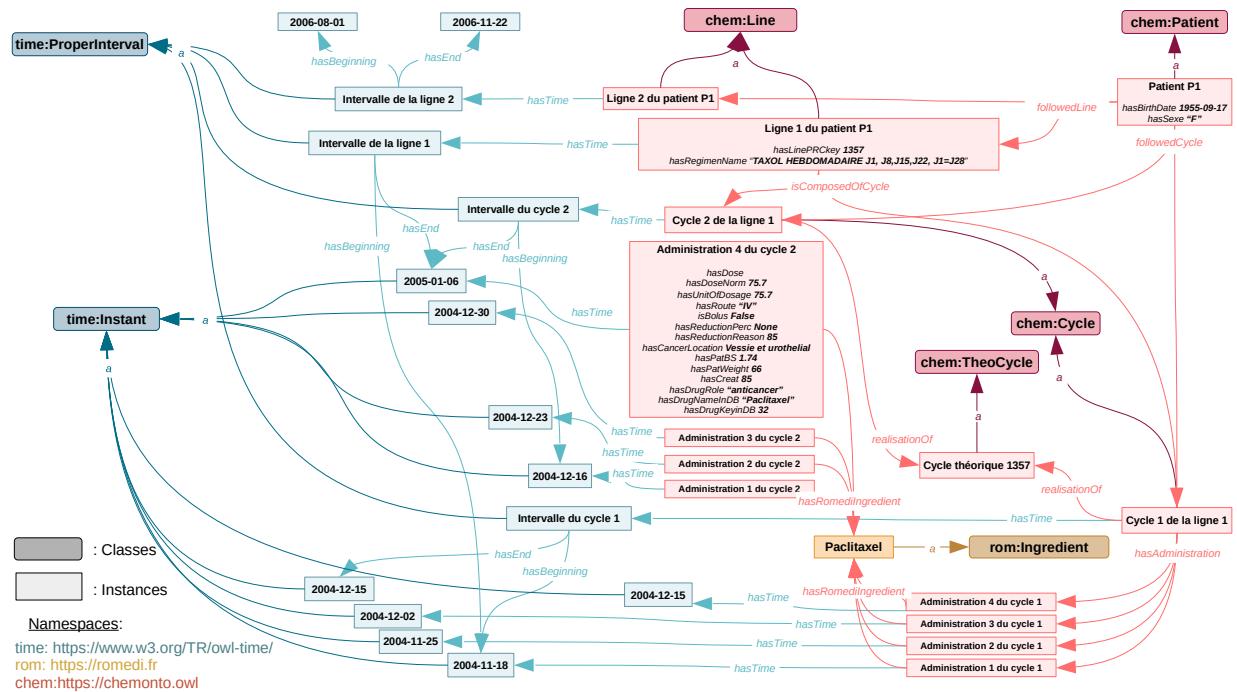


FIGURE 4.8 – Représentation ontologique de la première ligne du patient P1 suivant le schéma thérapeutique 1357 avec ChemoOnto. Pour alléger la visualisation, les propriétés de données associées sont indiquées en italique directement dans les instance, plutôt qu’avec des flèches, et elles ne sont indiquées que dans une seule des huit instances de "DrugAdministration"

CHAPITRE 4

Identifiant du patient	Clé du schéma thérapeutique	Nom du schéma thérapeutique	Numéro du cycle	Date de l'administration	Localisation de la tumeur	Identifiant Romedi	Ingrédient	Dose normalisée	Unité	Date de naissance du patient	Sexe du patient	Surface corporelle	Taille	Poids	Créatinine	Durée d'administration en minutes	Voie d'administration	Date de début de ligne	Date de fin de ligne	Date de début de cycle	Date de fin de cycle
P1	1357	TAXOL HEBDOMADAIRE J1, J8, J15, J22 J1-J28	1	2004-11-18	VESSIE et UROTHELIAL	Paclitaxel	INck2vkfdamu5ebj8j62fjq20pk8ltio	150.0	mg/m ²	1955-09-17	F	1.74	158	66	116	60	Voie Intra Veineuse	None	None	2004-11-18	2005-01-06
			1	2004-11-25	VESSIE et UROTHELIAL	Paclitaxel	INck2vkfdamu5ebj8j62fjq20pk8ltio	75.86	mg/m ²			1.74	158	66	116	60	Voie Intra Veineuse	None	None	2004-11-18	2005-01-06
			1	2004-12-02	VESSIE et UROTHELIAL	Paclitaxel	INck2vkfdamu5ebj8j62fjq20pk8ltio	75.86	mg/m ²			1.74	158	66	116	60	Voie Intra Veineuse	None	None	2004-11-18	2005-01-06
			1	2004-12-09	VESSIE et UROTHELIAL	Paclitaxel	INck2vkfdamu5ebj8j62fjq20pk8ltio	75.86	mg/m ²			1.74	158	66	116	60	Voie Intra Veineuse	None	None	2004-11-18	2005-01-06
			2	2004-12-16	VESSIE et UROTHELIAL	Paclitaxel	INck2vkfdamu5ebj8j62fjq20pk8ltio	75.86	mg/m ²			1.74	158	66	116	60	Voie Intra Veineuse	None	None	2004-11-18	2005-01-06
			2	2004-12-23	VESSIE et UROTHELIAL	Paclitaxel	INck2vkfdamu5ebj8j62fjq20pk8ltio	75.86	mg/m ²			1.74	158	66	85	60	Voie Intra Veineuse	None	None		
			2	2004-12-30	VESSIE et UROTHELIAL	Paclitaxel	INck2vkfdamu5ebj8j62fjq20pk8ltio	75.86	mg/m ²			1.74	158	66	85	60	Voie Intra Veineuse	None	None		
			2	2005-01-06	VESSIE et UROTHELIAL	Paclitaxel	INck2vkfdamu5ebj8j62fjq20pk8ltio	75.86	mg/m ²			1.74	158	66	85	60	Voie Intra Veineuse	None	None		
	1376	M-VAC HD (Cisplatin)	1	2006-08-01	VESSIE et UROTHELIAL	Cisplatin	INu1ijcoqh456snv1bvbtfs3ns4ge6qq	79.31	mg/m ²			1.74	158	66	85	15	Voie Intra Veineuse	None	None	2006-08-01	2006-11-22
			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
P2	1000	CISPLATINE 40mg/m ² J1-J2 - ALIMTA 500mg/m ² HDJ	1	2011-03-09	RESPIRATOIRE et THORACIQUE	Pemetrexed	INp5bj78bd6jmtcoin42tctshbo4old4aq9	483.87,0	mg/m ²	1951-04-03	M	1.86	162	74	60	10	Voie Intra Veineuse	None	None	2008-10-01	2009-06-08
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

FIGURE 4.9 – Représentation tabulaire de la première ligne du patient P1 suivant le schéma thérapeutique 1357 avec ChemoOnto. Les "|" indiquent une valeur identique à la dernière valeur observée de la colonne, tandis que les points "..." indiquent que les résultats continuent.

4.2.2 Extraire les connaissances sur la survenue de toxicités

Nous avons extrait les toxicités depuis deux sources de données dans l'entrepôt de l'HEGP : les questionnaire et les comptes rendus. Dans les comptes rendus, les toxicités sont présentes à la fois dans le texte libre et dans des tableaux. Nous avons donc utilisé trois processus d'extraction différents. Le projet github [OntoTox](#) permet d'instancier OntoTox avec les données provenant de différentes sources de cinq faux patients .

4.2.2.1 Reconnaissance des entités d'intérêt : toxicité et sévérité

Nous avons identifié deux entités d'intérêt : les toxicités et leur sévérité. La méthode de reconnaissance de ces entités est la même dans les trois sources. Nous avons créé un dictionnaire de toxicités basé sur trois terminologies sur les effets indésirables : la cinquième version du *Common Terminology Criteria for Adverse Events* (CTCAE) et la terminologie *World Health Organization Terminology*(WHOART) (*cf.* sections 1.2.1.2 et 1.2.3 du chapitre 1 et 2.3.4.3 du chapitre 2). Notre dictionnaire est constitué de 4 038 termes couverts par 835 concepts UMLS. Nous avons utilisé QuickUMLS pour extraire les toxicités [192] (*cf.* section 3.1.1.1). QuickUMLS est un module python qui utilise la méthode Simstring pour effectuer une correspondance approximative des chaînes de caractères. QuickUMLS a été appliqué en utilisant un critère de recouvrement de longueur, une similarité basée sur la distance de Jaccard et un seuil de 0,9. En plus de l'identification de la toxicité, l'approche basée sur le dictionnaire nous a également fourni l'identifiant unique de concept UMLS (appelé

CUI pour *Concept Unique identifier*), permettant la normalisation des toxicités. Nous avons utilisé une expression régulière pour détecter les grades et les normaliser en fonction de leur valeur numérique.

4.2.2.2 Lier les entités d'intérêt dans les différentes sources

Dans les questionnaires Nous avons identifié deux types de questionnaires qui relèvent les toxicités survenues lors des chimiothérapie. Ces questionnaires sont composés d'items correspondant à diverses toxicités et leurs réponses correspondent au grade de toxicité observé chez le patient. Il est trivial dans ce cas particulier de lier la toxicité à son grade, car les paires question-réponse des questionnaires sont structurées dans l'entrepôt.

Dans les comptes rendus, nous avons utilisé deux processus d'extraction différents pour extraire les toxicités et leurs sévérités associées dans les tableaux et le texte libre.

Dans le texte libre des comptes rendus cliniques Dans le texte libre, nous avons utilisé les parseurs de détection de contexte de PyMedExt [52] pour identifier si les toxicités extraites étaient dans une phrase négative ou hypothétique, et si elles concernaient le patient ou sa famille. PyMedExt est l'ancêtre de MedKit [86], une librairie python développée par l'équipe HeKA qui facilite l'extraction d'informations depuis diverses sources de données médicales, et notamment le texte libre. Nous avons intégré à PyMedExt, un parseur de relation entre entités d'intérêt, basé sur le parseur de dépendances syntaxique Stanza [163]. Stanza est une bibliothèque d'analyse du langage naturel en Python qui utilise le formalisme des dépendances universelles. Dans notre cas, l'objectif était d'identifier les liens entre les entités de toxicité et de grade. Nous avons traité toutes les phrases contenant au moins une entité de toxicité. Le parseur nous fournit un graphe de dépendances syntaxiques pour chaque phrase. Nous avons sélectionné de manière récursive toutes les entités qui se trouvaient sous la tête de l'entité de toxicité. Nous avons lié la toxicité et le grade s'il existait un chemin entre les deux entités. La figure 4.10 fournit un exemple d'extraction de relation entre les entités d'intérêt toxicité et grade avec un parseur de dépendance syntaxique.

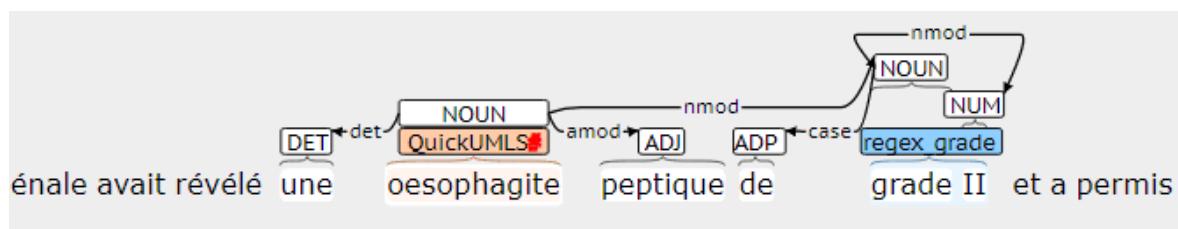


FIGURE 4.10 – Lier les entités toxicité et grade grâce au parseur de dépendance. L'entité toxicité est surlignée en orange et l'entité grade est surlignée en bleue

Dans les tableaux des compte rendus cliniques Pour distinguer les tableaux du texte libre, nous avons converti les documents au format html. Nous avons ensuite identifié et sélectionné les tableaux qui concernaient des toxicités et leurs grades associés en recherchant les termes "effets indésirables", "grade", "lié au traitement", "date de début" et "date de fin" dans les entêtes. Ces termes sont présents dans les templates word des documents du

département d'oncologie. Comme pour les questionnaires, il était trivial de lier toxicité et grade associé en se basant sur l'entête de la colonne dans laquelle l'entité d'intérêt avait été repérée.

La figure 4.11 schématisé les processus d'extraction de toxicités à partir des différentes sources et leur instanciation dans OntoTox.

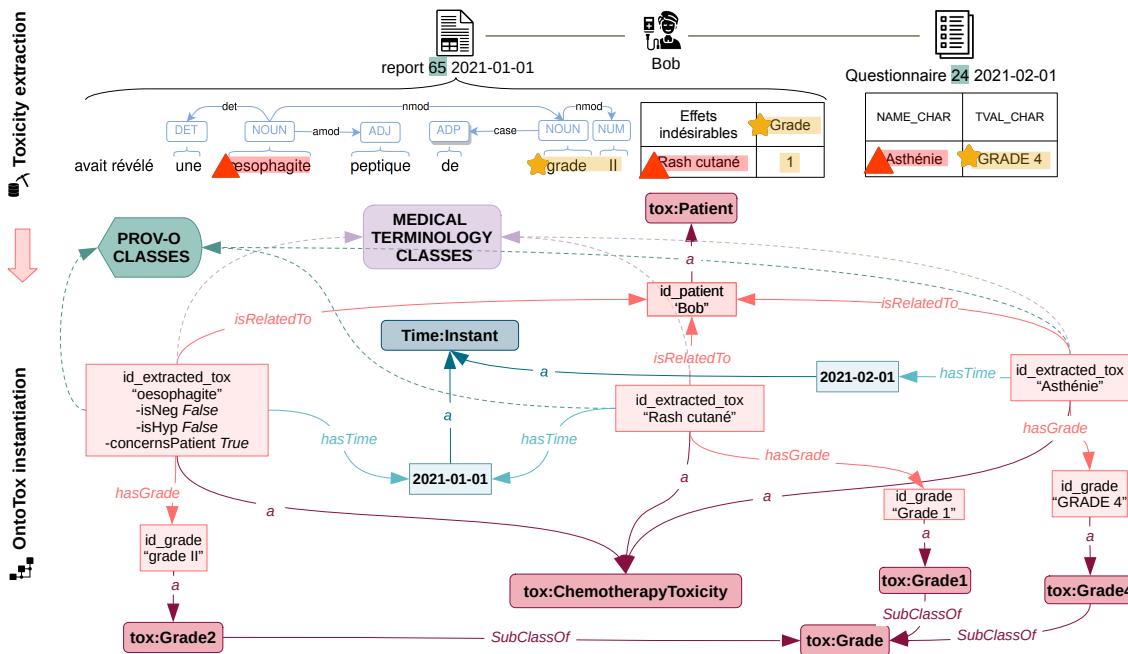


FIGURE 4.11 – De l'extraction des toxicités à partir de de différentes sources de données à leurs instantiation dans OntoTox

4.2.3 Discussion sur les méthodes d'extraction d'informations

4.2.3.1 Utilisation d'outils de Traitement Automatique des Langues (TAL) plus sophistiqués pour OntoTox

Nous sommes conscients que les outils de TAL utilisés tant pour reconnaître les entités d'intérêt (approches basées sur des dictionnaires pour les toxicités et expressions régulières pour les grades), que pour identifier les liens entre ces entités (parseur de dépendance) sont des méthodes assez simples et datées. Il faut noter que notre objectif premier n'était pas d'atteindre les meilleures performances d'extraction et de prédiction d'entités, mais plutôt de montrer la possible intégration des extractions dans un modèle commun. De plus, le développement de ces méthodes d'extraction (mai 2021) est antérieur à l'explosion des modèles de langages modernes et à leur facilité d'implémentation.

4.2.3.2 Évaluation de la qualité de l'extraction

OntoTox Bien que notre approche basée sur la correspondance de dictionnaires et les expressions régulières soit simple, nous pensons que l'utilisation de parseurs de dépendance aide à désambiguer l'extraction des grades. En oncologie, les grades peuvent également qualifier les stades tumoraux, et nous pensons avoir évité de tels faux positifs. Cependant ce ne sont que des suppositions, et notre travail d'extraction de toxicités et grades associés manque d'évaluation. Les reconnaissances d'entités avec le dictionnaire de toxicité et l'expression régulière pour identifier les grades n'ont pas non plus été évalués. L'évaluation de nos reconnaissances d'entités et de relations nécessiterait une annotation par plusieurs experts médicaux sur un corpus de documents volumineux avec un outil d'aide à l'annotation comme Brat [194] ou Doccano [88]. Une initiative d'évaluation a été lancée en septembre 2021, lors de l'intégration de l'annotateur de relation PyMedExt à MedKit (*cf.* section 4.2.2.2.2), mais nous ne l'avons malheureusement pas poursuivie.

ChemoOnto Comme nous l'avons évoqué dans le Préambule, les défis d'extraction d'informations se posent moins avec les données structurées des enregistrements du logiciel Chimio. À vrai dire, l'évaluation potentielle ici, ne consiste pas en une évaluation de la méthode d'extraction, mais plutôt en l'évaluation de la validité et la véracité des données qui sont enregistrées, à la fois pour les schémas thérapeutiques et pour les administrations réelles. Il faudrait faire vérifier, au moins pour un échantillon, la validité des schémas thérapeutiques enregistrés dans la base de données de Chimio par un groupe de pharmaciens. Pour les administrations réelles, il est difficile de corriger les potentielles erreurs d'enregistrements. Néanmoins, nous avons effectué une évaluation partielle en comparant, lorsqu'ils étaient renseignés, les pourcentages de réduction saisis par les soignant dans le champs "réduction" du logiciel Chimio (*cf.* section 1.4.6), accessibles dans la propriété de donnée "hasReductionPerc" de ChemoOnto (*cf.* 4.2.1.2), et la réduction calculée avec les valeurs des administrations théoriques et réelles. (*cf.* figure 4.12). Globalement, la comparaison est linéaire, ce qui augmente la confiance dans la qualité des données enregistrées pour les administrations réelles.

D'autre part, les pharmaciens de l'HEGP nous ont averti qu'il ne fallait pas se fier aux enregistrements des médicaments en prises en orale dans le logiciel, car ceux-ci ne sont pas systématiquement saisis par les soignants. Il faut donc accorder une confiance modérée à la qualité des schémas thérapeutiques contenant au moins une prise orale suivis par les patients.

4.2.3.3 Enrichissement progressif des ontologies

Amélioration de la qualification des toxicités extraites et de leur temporalité pour OntoTox Nous nous sommes concentrés sur l'extraction et l'intégration du type et du grade des toxicités. Une qualification plus détaillée de la toxicité, comme son traitement associé et sa durée, serait nécessaire pour offrir une vue plus complète des informations disponibles. Nous avons mentionné dans la section 4.2.3.2 que les dates associées aux toxicités manquent de précision car ce sont des métadonnées, elles concernent la date de la source dont la toxicité est extraite et non la date d'occurrence de la toxicité elle-même. De plus une toxicité est rarement ponctuelle, et il faudrait extraire sa durée associée. Dans le texte libre des comptes rendus, cela nécessite d'appliquer d'autres outils de TAL pour annoter la temporalité associée aux toxicités, comme par exemple l'annotateur SUTime [123].

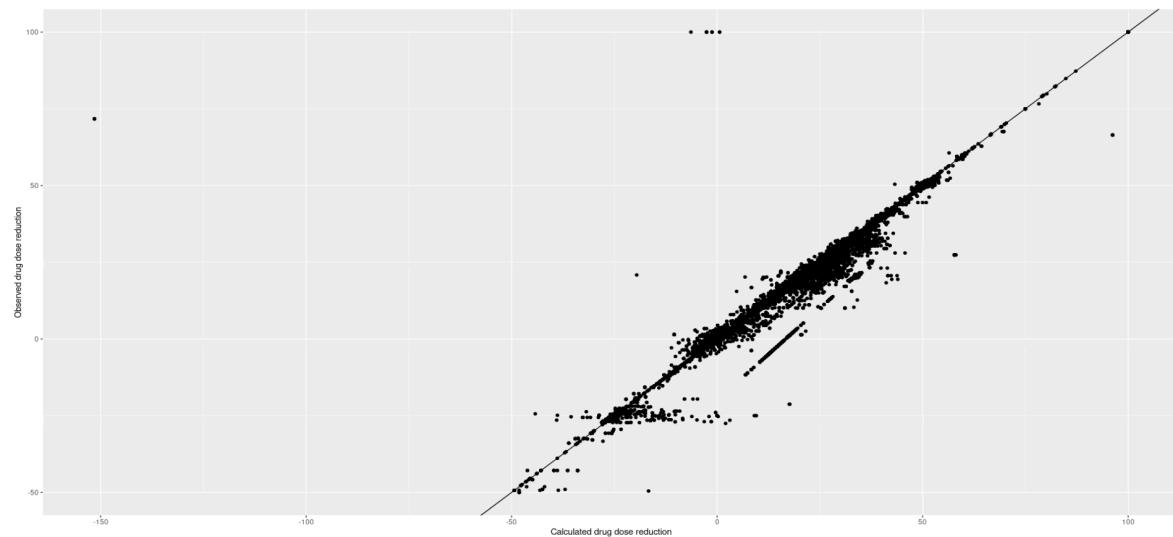


FIGURE 4.12 – Comparaison entre le pourcentage de réduction enregistré dans le champ "réduction" du logiciel Chimio, et le pourcentage de réduction calculée avec les valeurs des administrations théoriques et réelles.

Pour reconstruire les intervalles des toxicités à partir des questionnaires et des tableaux des compte rendus, on pourrait développer une stratégie semblable à celle adoptée pour reconstruire les cycles et les lignes de chimiothérapie pour ChemoOnto : ordonner les sources selon leur dates, sélectionner la première et la dernière mention de la toxicité et déduire les intervalles. Éventuellement pour les tableaux, si les champs "date de début" et "date de fin" sont remplis, il faudrait prioriser l'intervalle mentionné dans le tableau plutôt que l'intervalle déduit à partir des données. Enfin, il faudrait peut-être adopter, pour toutes les sources, des stratégies différentes en fonction du type de toxicité extraite. Par exemple, il serait sans doute plus efficace d'adopter une stratégie d'annotation différente, voire une représentation temporelle différente pour la temporalité des toxicités irréversibles (*cf. section 1.2.1.2*).

De nouvelles sources d'informations pour ChemoOnto et OntoTox Dans le logiciel Chimio, le champ "motif de la réduction", extrait dans ChemoOnto sous la propriété de données "hasReductionReason" est généralement rempli avec le nom d'une toxicité. De plus, les administrations d'anti-effets indésirables apportent indirectement une information sur les occurrences de toxicités. Ainsi, les enregistrements du logiciel Chimio pourraient constituer une source d'extraction de toxicités supplémentaire dans OntoTox. De même, les comptes rendus pourraient constituer une deuxième source d'informations pour reconstruire le cours des chimiothérapies dans ChemoOnto. On pourrait imaginer évaluer la reconstruction des intervalles de cycles et de lignes construits avec les enregistrements du logiciel Chimio en les comparant avec leur reconstruction à partir des comptes rendus.

Vers un score de confiance Enfin, il serait intéressant de développer un score de confiance sur l'information extraite et représentée qui dépendrait :

- de l'intersection de cette information dans différentes sources,

- de la méthode d'extraction,
- de la provenance de l'information extraite.

4.3 Utiliser les ontologies instanciées comme des bases de connaissances

4.3.1 Instanciation de ChemoOnto

Nous avons utilisé ChemoOnto pour représenter les schémas thérapeutiques théoriques et les lignes réellement suivies de chimiothérapies à partir des enregistrements du logiciel Chimio dans deux hôpitaux, l'HEGP et le CHU de Bordeaux. Pour les chimiothérapies réellement suivies, nous avons sélectionné les enregistrements sur une période s'étendant de juin 2002 à décembre 2021. La table 4.4 résume les résultats obtenus en comptabilisant le nombre d'instances de chaque classe de ChemoOnto. 41 930 lignes de 14 662 patients et 71 417 lignes de 33 725 patients ont respectivement été instanciées à l'HEGP et au CHU de Bordeaux. Pour les schémas thérapeutiques, on constate que le ratio du nombre d'administrations théoriques (nombre d'instances de la classe "TheoDrugAdministration") sur le nombre de schémas thérapeutiques (nombre d'instances de la classe TheoCycle) est plus élevé au CHU de Bordeaux ($\simeq 8$) qu'à l'HEGP ($\simeq 3$). Ceci s'explique peut-être par les différentes populations de patients traitées à l'HEGP et au CHU de Bordeaux.

Le projet ProtoDrift, détaillé dans le chapitre suivant (chapitre 5) a été développé en requérant le graphe ChemoOnto avec un ensemble de requêtes SPARQL, disponible dans le dossier `sparql_query_files` du projet gitlab `ProtoDrift` (disponible sur <https://gitlab.inria.fr/arogier/protodrift>). De plus, dans le projet gitlab `ChemoOntoTox` (disponible sur https://gitlab.inria.fr/arogier/ChemoOntoTox/-/tree/master?ref_type=heads), le dossier `SPARQL_query_ChemoOnto` contient une explication pour requérir le graphe sur les données de cinq faux patients, soit directement sur l'interface graphique de GraphDB, ou via l'API, avec le notebook `query_ChemoOnto.ipynb`.

Les différents résultats présentés dans cette section montrent l'utilisation de ChemoOnto comme base de connaissances, et son potentiel dans l'analyse et l'exploration du cours des chimiothérapies.

Classes de ChemoOnto	HEGP	CHU de Bordeaux
Patient	14 662	33 725
Line	41 930	71 417
Cycle	194 042	289 450
DrugAdministration	493 381	791 294
TheoCycle	1 358	7 083
TheoDrugAdministration	4 402	60 874

TABLE 4.4 – Répartition des instances des classes de ChemoOnto à l'HEGP et au CHU de Bordeaux

4.3.2 Instanciation d'OntoTox

OntoTox est une ontologie conçue pour représenter les toxicités liées aux chimiothérapies, leur sévérité et leur provenance. L'instanciation d'OntoTox vise à unifier les informations extraites de différentes sources hétérogènes afin de faciliter l'intégration et l'analyse des données relatives aux toxicités des traitements de chimiothérapie.

Pour montrer l'usage d'OntoTox comme base de connaissances, nous l'avons instanciée avec les données de l'entrepôt de l'HEGP, sur une cohorte de patients atteints d'un cancer du poumon. Ces patients ont été identifiés en utilisant le code de la Classification Internationale des Maladies (CIM10) C34 ("Néoplasme malin des bronches et du poumon") et 3 239 patients traités pour des cancers pulmonaires ont été sélectionnés. Parmi ces patients, nous en avons identifié 470 ayant au moins un compte rendu et un questionnaire sur les toxicités de chimiothérapie. Ensuite, nous avons sélectionné aléatoirement 330 patients pour constituer notre cohorte étudiée, en laissant de côté 140 patients pour une évaluation future. Pour chaque patient, tous les comptes rendus et les questionnaires de toxicité ont été extraits, avec un total de 11 819 comptes rendus et 71 140 items de questionnaire (cf. figure 4.13).

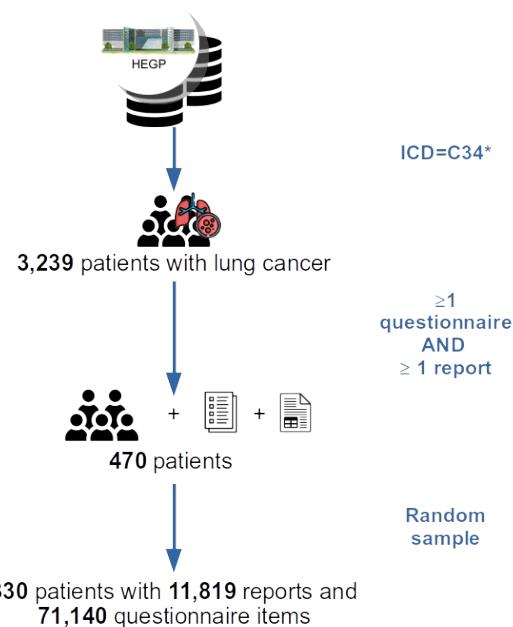


FIGURE 4.13 – Instanciation d'OntoTox avec une cohorte de patients atteints d'un cancer du poumon.

Nous avons instancié OntoTox avec 53,510 toxicités extraites des questionnaires, 54,420 extraites des textes libres et 2,366 extraites des tableaux. Cette instantiation révèle la complémentarité et la redondance des informations issues de ces trois sources. On constate que les toxicités extraites des questionnaires sont systématiquement associées à une absence explicite de la toxicité en question (77%), ou à un grade non nul (33%), ce qui souligne leur nature structurée. Par contraste, la grande majorité (89%) des toxicités extraites des textes

Classes d'OntoTox	Questionnaires	Texte libre	Tableaux
ChemotherapyToxicity	53 510	54 420	2 366
Grade	53 510	6 366	400
Grade1	9 981	2 100	87
Grade2	1 832	1 996	52
Grade3	191	817	23
Grade4	19	422	0
Grade5	0	2	1
GradeNull	0	433	85
Grade0	41 487	596	152
Patient	330	330	330
StartDate	1 112	2 782	372

TABLE 4.5 – Répartition des instances des classes d'OntoTox extraites de trois sources différentes : les questionnaires cliniques, les textes libres et les tables semi-structurées. Cette table montre le nombre total de toxicités et de grades détectés dans chaque source. On rappelle que "GradeNull" fait référence à un grade dont on n'a pas pu détecter le numéro, tandis que "Grade0" fait référence à l'absence explicite de la toxicité.

libres n'est pas associée à un grade ou associée à un grade nul.

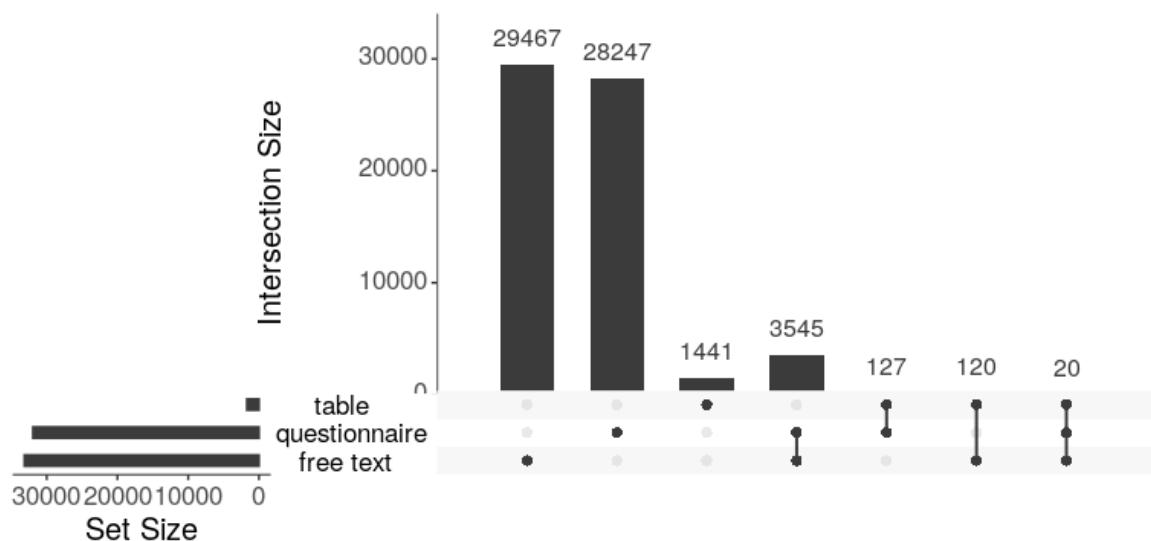


FIGURE 4.14 – Diagramme d'intersection des toxicités extraites depuis les différentes sources

Pour générer le diagramme d'intersection 4.14, nous avons interrogé OntoTox avec une requête SPARQL en sélectionnant les toxicités extraites normalisées avec leurs concepts UMLS, par patient et par mois de la date associée à l'extraction, et regroupées par sources.

La possibilité d'interroger le graphe via SPARQL démontre l'utilité d'OntoTox en tant que base de connaissances. On constate que les trois sources ne se chevauchent que modérément. Cette faible concordance souligne l'importance d'intégrer plusieurs sources pour obtenir une vue d'ensemble complète des toxicités.

Nous pensons que les problèmes de qualité d'extraction, particulièrement sur les dates associées aux toxicités (discutés dans la section 4.2.3.3), expliquent en partie la faible concordance entre sources. Les dates associées aux extractions de toxicités sont des méta-données, elles sont associées à l'édition de la source d'extraction (date d'édition du questionnaire ou du compte rendu), qui ne correspond pas forcément à la date de l'occurrence de la toxicité extraite. Ces dates manquent donc de précision. À titre d'exemple, l'un des questionnaires analysés est utilisé par les soignants pour collecter les toxicités rencontrées par le patient lors de son dernier cycle, la veille du cycle suivant. Il y a donc souvent un écart de dates entre les occurrences des toxicités collectées et la date de collecte, attribuée au questionnaire. À l'échelle du traitement de chimiothérapie, cet écart a son importance.

De plus, la normalisation des toxicités a été effectuée au niveau du Concept Unique Identifier (CUI) de l'UMLS. Or, cette normalisation est parfois trop précise, et il serait plus pertinent de tirer parti de la structure hiérarchisée des concepts de l'UMLS pour adapter le niveau de normalisation à la toxicité extraite. Par exemple, "souffle court" (CUI : C0013404) et "difficulté à respirer" (CUI : C0013428) sont comptés comme deux toxicités distinctes. Dans ce cas, il serait plus juste de normaliser ces deux concepts avec leur parent commun afin de les comptabiliser comme une intersection si elles sont rencontrées dans diverses sources.

Malgré ces réserves, OntoTox a prouvé son potentiel à unifier et intégrer les informations sur les toxicités de chimiothérapie provenant de diverses sources. Cette intégration enrichit la base de connaissances et facilite l'exploration et l'analyse des données. L'instanciation d'OntoTox souligne la pertinence d'utiliser des ontologies pour créer des bases de connaissances robustes et exploitables.

4.4 Utiliser les ontologies pour découvrir de nouvelles connaissances

4.4.1 Raisonnement temporel

Nous avons instancié ChemoOnto et OntoTox sur une cohorte de 3 923 patients ayant au moins un enregistrement d'administration anticancéreuse et un questionnaire de toxicité enregistré dans l'entre�ot de données de l'HEGP (*cf. figure 4.15*). Dans ce graphe, nous avons harmonisé l'espace de nommage des instances de la classe Patient pour qu'il soit commun aux deux ontologies. Le résultat est une base de connaissances intégrant à la fois des toxicités instanciées dans OntoTox et le déroulement des chimiothérapies instanciées dans ChemoOnto. Pour ces 3 923 patients, 220 483 toxicités avec leurs grades de sévérité ont été instanciées dans OntoTox, et 11 197 lignes de chimiothérapies ont été instanciées dans ChemoOnto (*cf. tableaux 4.6 et 4.7*).

Classes d'OntoTox	# Instances	Classes de ChemoOnto	# Instances
ChemotherapyToxicity	980 105	Administration	487 325
Toxicité évaluée comme absente	759 622	Cycle	62 343
Toxicité évaluée avec sa sévérité	220 483	Line	11 197

TABLE 4.6 – Classes d'OntoTox et nombres d'instances

TABLE 4.7 – Classes de ChemoOnto et nombres d'instances

Nous avons intégré à cette base de connaissances une règle SWRL (*cf.* figure 4.16), pour inférer la propriété **time:inside**, liant des instants associés à des toxicités à des intervalles associés à des cycles de chimiothérapies suivis. Cette propriété est l'équivalent de la relation de Allen **time:intervalDuring** entre deux intervalles présentée dans la sous-section 2.4.1 du chapitre 2, mais elle lie un instant à un intervalle. Le raisonneur Pellet [187] a été utilisé pour inférer cette propriété. Pour limiter le temps computationnel, nous avons divisé la base de connaissances en fichiers de 500 patients et parallélisé le raisonnement sur quatre processeurs.



FIGURE 4.15 – Instanciation de ChemoOnto et OntoTox pour 3 923 patients avec des items de questionnaires et des enregistrements dans le logiciel Chimio

 FIGURE 4.16 – Règle SWRL pour inférer la propriété **time:inside** entre les instances d'instants liés à des toxicités et les instances d'intervalles liés à des cycles suivis

La table 4.8 montre les résultats des inférences. 191 436 toxicités avec leur sévérité ont été liées à des cycles suivis via le raisonnement temporel. Ce résultat démontre le potentiel des ontologies à lier des bases de connaissances via la temporalité. Toutefois, ce potentiel

est limité par les capacités de calcul des raisonneurs (*cf.* sous-section 2.4.1.1), et le fait que le raisonnement temporel est un problème NP-complet (*cf.* sous-section 2.4.1). Cette complexité nous a conduit à limiter notre raisonnement à une seule relation spécifique avec la règle SWRL (*cf.* figure 4.16), et à restreindre les entités temporelles ciblées.

Propriété d'objet time :inside	# Instances
Toxicité instant identifiée dans un intervalle de cycle	825,202
Toxicité évaluée comme absente instant identifiée dans un intervalle de cycle	633,766
Toxicité évaluée avec sa sévérité instant identifiée dans un intervalle de cycle	191,436

TABLE 4.8 – Instances de la propriété d'objet time :inside

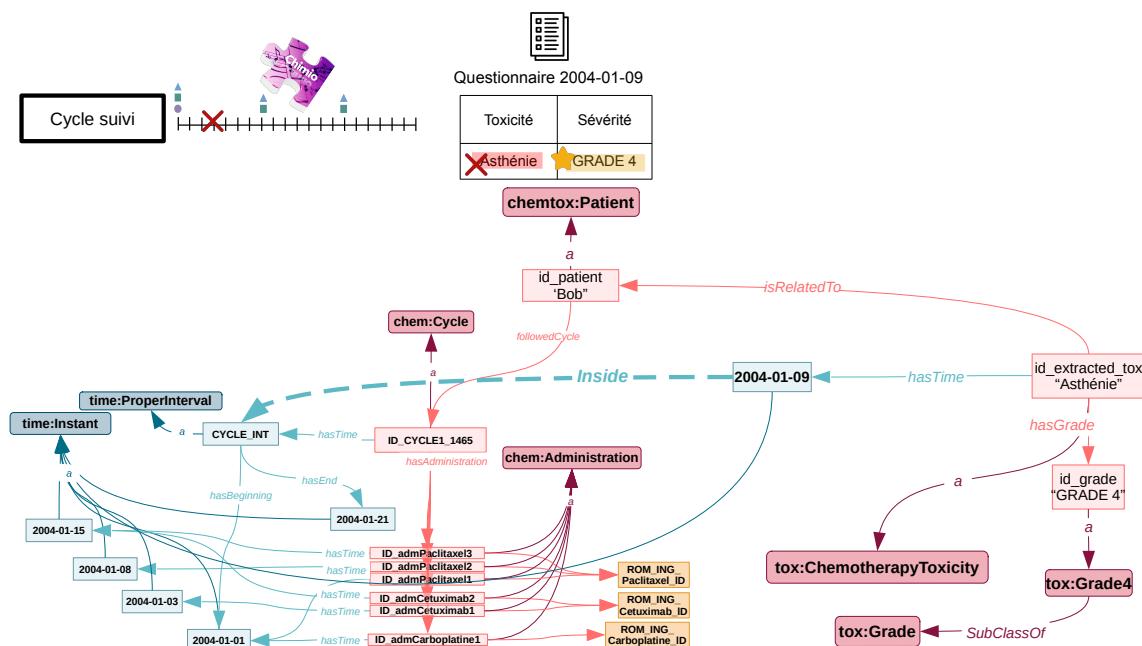


FIGURE 4.17 – Inférer la propriété `time:inside` entre les instances d'instants liés à des toxicités et les instances d'intervales liés à des cycles suivis

Notre règle spécifie que les toxicités et les cycles suivis doivent être liés au même patient, limitant ainsi les relations non pertinentes entre des entités temporelles de patients différents. De plus, nous spécifions quelles entités temporelles (les instants liés aux toxicités et les intervalles liés aux cycles suivis) doivent être inférées avec la propriété `time:inside`. Par conséquent, les autres relations de l'algèbre d'Allen, comme `time:contains` (inverse de `time:inside` ou `time:intervalDuring`) ou `time:before`, ne sont pas inférées.

Dans le contexte de notre étude, il n'est pas nécessaire d'inférer toutes les relations simultanément. Il serait intéressant de pouvoir spécifier, au sein d'un triplestore tel que GraphDB, des règles SWRL indiquant quelles relations temporelles inférer au moment de

la requête SPARQL. Ainsi, la base de connaissances ne contiendrait aucune relation de l’algèbre d’Allen par défaut, mais celles-ci seraient inférées en fonction des besoins de l’utilisateur au moment du requêtage. À notre connaissance, cela n’est pas encore possible, mais ce serait particulièrement utile pour lier des bases de connaissances contenant des données cliniques temporelles.

4.4.2 Comparer les schémas thérapeutiques avec ChemoKG et les plongements de graphes

Nous présentons ici une réutilisation de ChemoOnto pour étudier et comparer les schémas thérapeutiques de chimiothérapies avec des méthodes de plongements de graphes [105].

La partie sur les schémas thérapeutiques de ChemoOnto a été intégrée au sein du graphe de connaissances ChemoKG. ChemoKG lie les médicaments anti-cancéreux aux composés chimiques de l’ontologie ChEBI (*cf.* table 2.1), ce qui enrichit le graphes d’informations telles que la demi-vie des molécules. L’ensemble forme la base de connaissance ChemoKG, requérable via le point d’entrée <http://chemokg.paris.inria.fr/sparql>.

Le graphe ChemoKG a été exploré avec cinq algorithmes de plongements de graphes : TransE, DisMult, MuRE, ComplEx et CompGCN. En plus de ces méthodes, un nouvel algorithme a été développé par Jong Ho Jhee, inspiré de Réseaux Neuronaux de Graphes (Graphs Neural Networks - GNN), nommé "Relationale Graph Embedding" (RAGE). Avec RAGE, chaque entité est encodée avec un vecteur d’embedding qui correspond à une aggrégation de l’ensemble des informations présente dans ses entités voisines.

Pour évaluer le nouvel algorithme de plongement de graphe RAGE, deux tâches ont été réalisées :

- Prédiction de lien : Évaluation des performances de RAGE et des autres algorithmes sur une tâche de prédiction de liens. Les résultats montrent que RAGE surpassé quatre des cinq algorithmes de référence.
- Clustering des schémas thérapeutiques : Clustering des embeddings des schémas thérapeutiques obtenus avec RAGE et MuRE. Les résultats montrent que RAGE permet un meilleur regroupement des protocoles selon les localisations de cancer et les classes ATC des médicaments.

Les résultats montrent que RAGE est capable de capturer efficacement les relations entre les entités dans ChemoKG, ce qui permet de regrouper les schémas thérapeutiques de chimiothérapie de manière cohérente. Cette étude met en lumière le potentiel des graphes de connaissances et des techniques de plongements de graphes pour comparer et analyser les schémas thérapeutiques. Les plongements de graphes facilitent la mise en évidence des similitudes et des différences entre ces schémas, et offre ainsi une nouvelle perspective sur l’organisation et l’évaluation des traitements. Avec le clustering, on distingue d’une part des localisations avec des protocoles très spécifiques. D’autre part on distingue des groupes de protocoles très proches qui sont utilisés pour diverses localisations de cancer. Ces résultats suggèrent qu’une investigation clinique approfondie serait nécessaire pour mieux comprendre les regroupements obtenus et leur pertinence clinique.

4.5 Bilan

Ce chapitre a présenté deux contributions pour l'extraction et la représentation des réponses aux chimiothérapies à partir des données hospitalières : le développement des ontologies **ChemoOnto** et **OntoTox**. Ces ontologies ont démontré leur pertinence et leur efficacité pour structurer, intégrer et explorer des connaissances médicales complexes.

ChemoOnto permet une représentation précise des cycles et des lignes de chimiothérapie, en intégrant des données de la Time Ontology pour la temporalité et de Romedi pour la normalisation des médicaments. La structure de ChemoOnto facilite la comparaison entre les traitements théoriques et réels, offrant ainsi une base solide pour l'analyse des schémas thérapeutiques et l'évaluation des réponses aux traitements. (*cf. sections 4.1.1.1 et 4.1.2.1*)

OntoTox, de son côté, standardise et qualifie les toxicités associées aux chimiothérapies en utilisant des terminologies de référence telles que UMLS et MedDRA. L'intégration de l'ontologie PROV-O assure la traçabilité des données extraites. (*cf. sections 4.1.1.2 et 4.1.2.2*)

Les méthodes d'extraction développées pour ces ontologies ont permis de capturer des connaissances à partir de diverses sources de données, qu'elles soient structurées, semi-structurées ou non structurées. Bien que ces méthodes reposent sur des techniques relativement simples de reconnaissance d'entités et de parseur de dépendance, elles ont montré leur capacité à intégrer efficacement les extractions dans un modèle commun. Cependant, des améliorations sont nécessaires, notamment l'utilisation de modèles de traitement du langage naturel plus sophistiqués et une meilleure qualification de la temporalité des toxicités. (*cf. section 4.2*)

OntoTox a été instanciée avec les données de l'entrepôt de l'HEGP, sur une cohorte de patients atteints de cancer du poumon, ce qui a démontré sa capacité à unifier et structurer les informations sur les toxicités provenant de différentes sources (*cf. 4.3.2*). L'instanciation de ChemoOnto, quant à elle, a été réalisée avec succès dans deux hôpitaux, couvrant une période de près de vingt ans, ce qui témoigne de sa portabilité. Ces instances ont permis de créer une base de connaissances facilitant les analyses approfondies des traitements et de leurs effets. (*cf. section 4.3.1*)

De plus, l'exploration des connaissances avec le raisonnement temporel a permis de lier des toxicités spécifiques à des intervalles de cycles de chimiothérapie suivis, en utilisant des règles SWRL pour inférer des relations temporelles. L'utilisation d'algorithmes de plongement de graphes existants, ainsi que le nouvel algorithme RAGE, a permis de clusteriser les schémas thérapeutiques de manière cohérente. Ces techniques ont montré des résultats intéressants mais qui nécessitent des investigations approfondies. (*cf. section 4.4*)

En conclusion, les ontologies ChemoOnto et OntoTox, ainsi que les méthodes d'extraction développées, offrent des outils puissants pour la représentation et l'analyse des chimiothérapies et de leurs réponses. Les perspectives futures incluent l'amélioration des techniques d'extraction, le développement de scores de confiance pour les informations extraites, et la poursuite de l'exploration de connaissances avec le raisonnement et les plonge-

ments de graphes. Ces perspectives devraient enrichir les bases de connaissances actuelles.

ChemoOnto a facilité le développement de ProtoDrift, qui mesure l'adhésion aux traitements de chimiothérapies. Le chapitre suivant présente cette mesure.

CHAPITRE

5

PROTODRIFT : MESURER L'ADHÉSION AUX CHIMIOTHÉAPIES

5.1	Motivation pour une nouvelle mesure de l'adhésion aux chimiothérapies	130
5.1.1	Liens entre adhésion et réponses aux chimiothérapies	130
5.1.2	Hypothèses sur la mesure actuelle d'adhésion et objectifs de ProtoDrift	132
5.1.3	Formalisation des concepts de chimiothérapies	132
5.2	Définition de ProtoDrift	134
5.2.1	Dissimilitarité administration	134
5.2.2	Dissimilitarité médicament	136
5.2.3	Dissimilitarité intra-cycle et dissimilitarité inter-cycle	138
5.2.4	Dissimilitarités : du cycle à la ligne	139
5.2.5	Exemple de calcul de dissimilitarité	140
5.3	Optimisation des poids de ProtoDrift	142
5.3.1	Introduction à l'optimisation	142
5.3.2	Objectifs de l'optimisation	143
5.3.3	Méthode d'optimisation	144
5.4	Analyse comparative entre ProtoDrift et la RDI	151
5.4.1	Analyse du pouvoir discriminant	153
5.5	Exploration de l'impact du temps par rapport à la dose sur la survie globale	154
5.6	Bilan sur les méthodes proposées et leurs évaluations	155
5.7	Application de ProtoDrift sur les données l'HEGP et du CHU de Bordeaux	157
5.7.1	Sources de données et conception de l'étude	157
5.7.2	Correspondance entre les noms de localisation de cancer à l'HEGP et au CHU de Bordeaux	162
5.7.3	Évaluation de ProtoDrift sur deux jeux de données indépendants .	163
5.7.4	Configuration des paramètres d'application	163
5.8	Résultats d'application	164
5.8.1	Performances prédictives de la survie globale à 5 ans de la cohorte respiratoire et thoracique en première ligne	164
5.8.2	Évaluation de ProtoDrift	167
5.8.3	Optimisation des poids de ProtoDrift	171

5.8.4	Exploration de l'impact du temps par rapport à la dose sur la survie globale	172
5.9	Discussion et perspectives	173
5.9.1	Interprétation des résultats d'application	173
5.9.2	Optimisation des poids de ProtoDrift	177
5.9.3	Conception de la mesure ProtoDrift	180
5.10	Bilan	181
5.10.1	Résumé des contributions	181
5.10.2	Implications pour la recherche clinique	181
5.10.3	Défis et perspectives	182
5.10.4	Conclusion	182



Dans le chapitre 1, nous avons exposé les différentes stratégies médicales en cas de non réponse au traitement ou de survenue de toxicités. En fonction de l'état du patient, l'oncologue peut décider de modifier les jours, les doses d'administrations, les débuts de cycle ou même de changer de ligne de traitement. De plus, nous avons souligné le fait que les oncologues n'étaient pas toujours d'accord sur les décisions à prendre. Nous avons constaté dans la section 1.3 que de nombreuses études utilisaient la dose-intensité relative (RDI) pour prouver l'efficacité ou la non-efficacité d'ajustements de protocoles. Nous avons également mis en relief les limites de cette mesure.

Dans ce chapitre, nous présentons une dernière contribution : le développement de ProtoDrift, une métrique pour mesurer à différents niveaux une dérive au traitement sous chimiothérapie.

5.1 Motivation pour une nouvelle mesure de l'adhésion aux chimiothérapies

En s'appuyant sur les concepts et les problématiques évoquées au cours du chapitre 1, on motive ici le développement d'une nouvelle méthode pour mesurer l'adhésion aux chimiothérapies, ProtoDrift.

5.1.1 Liens entre adhésion et réponses aux chimiothérapies

Les deux premières parties (sections 1.1 et 1.2) du chapitre 1 ont mis en évidence la complexité à la fois théorique et pratique des chimiothérapies. Le traitement par chimiothérapie vise à trouver un équilibre optimal entre deux objectifs : minimiser les survenues de toxicités et minimiser la progression du cancer (*cf. section 1.2.1*). Pour atteindre cet équilibre, les schémas thérapeutiques sont méthodiquement définis (*cf. section 1.1.4*). Quand un cancer est diagnostiqué, une équipe d'experts médicaux se réunit et sélectionne rigoureusement un schéma thérapeutique (*cf. section 1.2.2*), en tenant compte du type, du stade et de la localisation du cancer et du profil du patient. Le traitement par chimiothérapie consiste en la répétition de cycles définis par le schéma thérapeutique sélectionné.

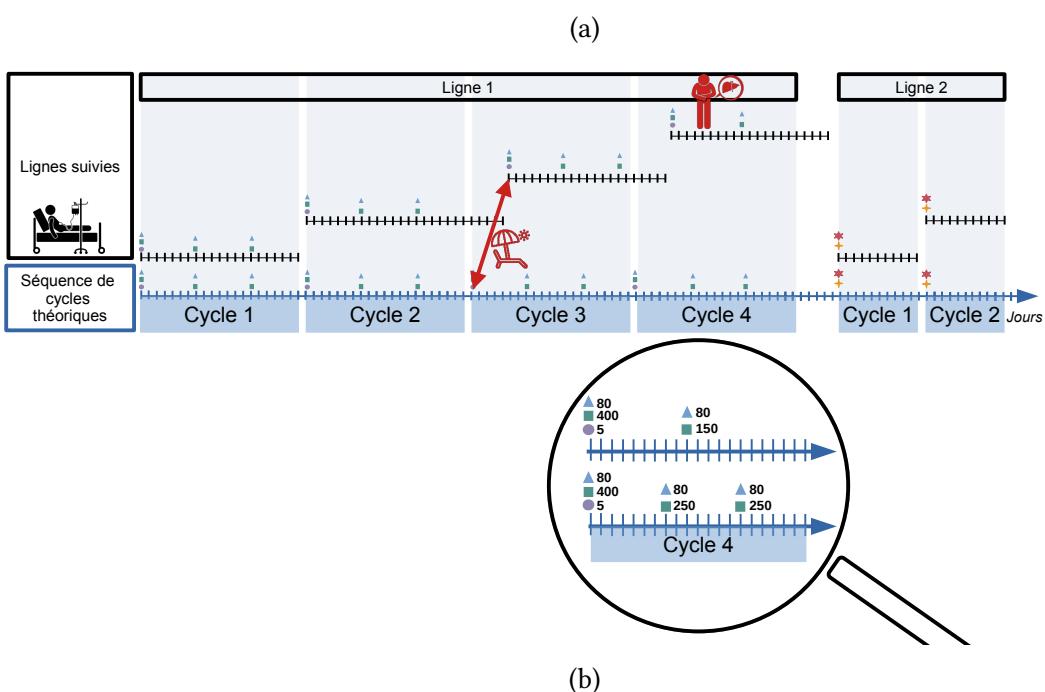
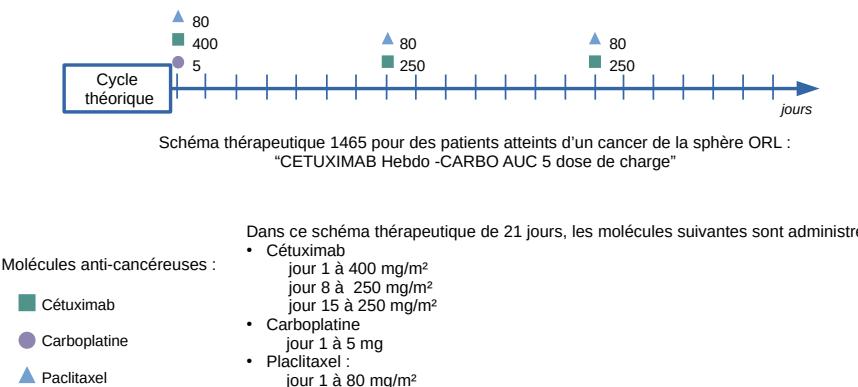


FIGURE 5.1 – (a) Description graphique du schéma thérapeutique 1465 utilisé pour traiter des cancers ORL. (b) Comparaison entre une ligne suivie par un patient fictif et le protocole (*i.e.* la séquence de cycles théoriques définis par le schéma 1465). Les lignes suivies sont représentées par les séquences en escalier noir, où chaque palier représente un cycle suivi. Le protocole de chimiothérapie, c'est-à-dire la séquence de cycles théoriques décrits par le schéma thérapeutique 1465, est représentée en bleue. Dans ce scénario, le cycle 3 est retardé car le patient est en vacances. Au cycle 4, l'oncologue décide de changer le jour d'administration des anticancéreux Cétuximab et Paclitaxel en raison d'une toxicité hépatique, et de réduire la dose de Cétuximab. À la fin du cycle 4, après avoir constaté que la tumeur ne diminue pas assez rapidement, l'équipe d'experts médicaux décide de changer de schéma thérapeutique. Ainsi, le patient entame une nouvelle ligne de traitement.

Malgré la précision de ces protocoles, leur mise en œuvre chez les patients subit souvent des modifications. Ces ajustements peuvent résulter de stratégies médicales visant à équilibrer efficacité et toxicité du traitement, d'adaptations pour le confort des patients, ou de contraintes opérationnelles au sein des établissements de santé. Ces ajustements sont établis dans le cadre d'une coopération continue entre patients et médecins. Le degré de ces écarts par rapport au traitement est qualifié d'adhésion thérapeutique. Un scénario d'adhésion au traitement par chimiothérapie est présenté dans la figure 5.1.1.

Les ajustements aux protocoles sont également motivés par de nombreuses hypothèses empiriques que les oncologues formulent à propos des protocoles de chimiothérapie, lesquelles varient considérablement en fonction du profil du patient. Ces adaptations ne font pas toujours consensus au sein de la communauté oncologique (*cf. section 1.2.6*). De plus, l'impact des modifications de calendrier, qui sont elles souvent causées par les préférences des patients, reste peu connu. Le manque de consensus et la méconnaissance des effets des écarts aux traitements sur la survie soulignent la nécessité d'outils capables de fournir une description précise de ces écarts et d'évaluer leurs répercussions sur l'évolution clinique des patients.

5.1.2 Hypothèses sur la mesure actuelle d'adhésion et objectifs de ProtoDrift

La RDI est actuellement la méthode standard pour évaluer l'adhésion aux protocoles de chimiothérapie (*cf. section 1.3.2*). Cependant, nous pensons que la RDI simplifie les dynamiques complexes des traitements de chimiothérapie, en se concentrant sur le dosage sans tenir pleinement compte du calendrier des administrations et de ses impacts potentiels sur les résultats des patients.

Nous proposons ProtoDrift, une approche plus nuancée pour quantifier les déviations par rapport au traitement prévu, en attribuant des dissimilarités pondérées à chaque type de variation. Cela permet ainsi de capturer à la fois les soins médicaux fournis et l'état du patient tout au long de la ligne de traitement. De plus, les poids de ProtoDrift peuvent être optimisés pour ajuster l'importance des différents facteurs en fonction du contexte. En offrant une vue détaillée des ajustements effectués pendant la chimiothérapie, ProtoDrift vise à améliorer notre compréhension des réponses au traitement et à guider une gestion plus efficace des patients. Il cherche à aller au-delà des limitations de la RDI pour une évaluation plus précise de l'adhésion à la chimiothérapie et de ses implications pour les soins contre le cancer.

Avant de présenter ProtoDrift, son optimisation et l'évaluation de son intérêt dans la section 5.2, nous définissons les notations que nous utiliserons dans la suite du chapitre.

5.1.3 Formalisation des concepts de chimiothérapies

Dans cette section nous formalisons un certain nombre de concepts sur les chimiothérapies nécessaires à la définition de ProtoDrift.

5.1.3.1 Notations des concepts de chimiothérapies

- un $i^{\text{ème}}$ cycle de chimiothérapie suivi est défini par un ensemble d'administrations $\mathcal{C}^i = \{a^1, \dots, a^n\}$. Chaque cycle a une durée $\tau_{\mathcal{C}^i}$. On notera \mathcal{C}^{theo} le cycle théorique défini par le schéma thérapeutique de durée $\tau_{\mathcal{C}^{theo}}$.
- une ligne \mathcal{L} est un ensemble ordonné de cycles suivis : $\mathcal{L} = \{\mathcal{C}^1, \dots, \mathcal{C}^l\}$. Une ligne a une durée $\tau_{\mathcal{L}}$.
- un protocole est une répétition de cycles théoriques.
- un médicament anti-cancéreux m est une paire (*molécule, mode d'administration*). Par exemple (cetuximab, perfusion continue) ou (cetuximad, bolus).
- une administration de médicament a est composée de trois dimensions : $\langle a_m, a_t, a_d \rangle$:
 - a_m est le médicament anti-cancéreux m associé à l'administration,
 - a_t est la durée entre le début du cycle et l'administration. Par exemple, 7 si l'administration doit être effectuée 7 jours après le début du cycle,
 - a_d est la dose. Par exemple 300 si l'administration doit être effectuée avec une dose de 300 mg.
- \mathcal{C}_m^i est le sous-ensemble des administrations du cycle \mathcal{C}^i qui ont le même médicament anti-cancéreux m (c'est-à-dire, la même paire (*molécule, mode d'administration*)).
- \mathcal{M} est l'ensemble des médicaments anti-cancéreux de \mathcal{C}^i ou \mathcal{C}^{theo} :

$$\mathcal{M} = \{m\} \cup \{m_{theo}\} \quad (5.1)$$

où $\{m\}$ (respectivement $\{m_{theo}\}$) est l'ensemble de médicaments du cycle \mathcal{C}^i (resp. \mathcal{C}^{theo}).

5.1.3.2 Notations de la dose-intensité relative (RDI)

On réécrit les différents composants de la RDI, définie dans le chapitre 1, avec les notations que l'on vient d'établir.

Pour un médicament anticancéreux m , une ligne \mathcal{L} et sa durée en semaines $\tau_{\mathcal{L}}$:

- La dose-intensité DI (*cf. formule 1.1*) est définie par :

$$\text{DI}(\mathcal{L}, m) = \frac{1}{\tau_{\mathcal{L}}} \sum_{i=1}^l \sum_{\forall a \in \mathcal{C}_m^i} a_d \quad (5.2)$$

- La dose-intensité théorique :

$$\text{DI}(\mathcal{C}^{theo}, m) = \frac{1}{\tau_{\mathcal{C}^{theo}}} \sum_{\forall a \in \mathcal{C}_m^{theo}} a_d \quad (5.3)$$

- La dose-intensité relative (RDI) (*cf. formule 1.2*) d'une ligne \mathcal{L} :

$$\text{RDI}(\mathcal{L}, \mathcal{C}^{theo}, m) = \frac{\text{DI}(\mathcal{L}, m)}{\text{DI}(\mathcal{C}^{theo}, m)} \quad (5.4)$$

- L'ADRDI (*All Drugs RDI*) pour tous les médicaments dans une ligne de traitement \mathcal{L} est définie comme suit :

$$\text{ADRDI}(\mathcal{L}, \mathcal{C}^{theo}) = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \text{RDI}(\mathcal{L}, \mathcal{C}^{theo}, i) \quad (5.5)$$

L'ADRDI fournit une mesure normalisée pour mesurer l'adhésion, en d'autres termes, pour comparer la ligne suivi au protocole prévu, avec des valeurs variant entre 0 et 1. Une valeur d'ADRDI plus élevée indique une bonne adhésion et une ligne suivie plus proche du protocole. Il s'agit là du contraire des dissimilarités de ProtoDrift que nous verrons dans la section 5.2, où une valeur plus élevée signifie une mauvaise adhésion et une déviation importante au traitement.

Notons une nouvelle fois que la RDI et l'ADRDI ne sont pas des distances, car elles ne vérifient pas la propriété de symétrie (*cf.* définitions 3.1.2.1). Ceci est directement à lier à leur objectif qui est de mesurer un écart par rapport à une référence, pour mesurer une adhésion thérapeutique. Ce sont des rapports normalisés.

Notation - Baseline All Drugs Relative Dose-Intensity (BADRDI)

Dans la suite du chapitre, on nommera la méthode de mesure de l'adhésion aux chimiothérapies avec la variable ADRDI, la méthode "BADRDI" pour *Baseline All Drugs Relative Dose-Intensity*.

5.1.3.3 Définition de la dérive aux chimiothérapies

Avec ProtoDrift, nous cherchons à mesurer la dérive par rapport au protocole de chimiothérapie, c'est-à-dire la non-adhésion au traitement de chimiothérapie. Pour calculer la dissimilarité entre le traitement théorique et le traitement réel, nous prenons en compte les paramètres suivants :

- Pour chaque administration au sein du cycle :
 - a_t^i : le temps en jours de depuis le début du cycle
 - a_d^i : la dose
- Au sein d'une ligne :
 - C_{start}^i : le temps en jours entre le début du cycle C^i et le début de la ligne

Dans la section suivante, nous montrons comment mesurer cette dérive au traitement avec ProtoDrift.

5.2 Définition de ProtoDrift

5.2.1 Dissimilité administration

À différentes échelles temporelles du traitement par chimiothérapie, nous définissons des dissimilarités traduisant les écarts par rapport au protocole prévu. Ces dissimilarités

sont définies entre 0 et 1 : lorsque la dissimilarité est proche de 0, cela signifie que le patient suit le protocole. Dans le cas où la dissimilarité vaut 1, le protocole n'est pas du tout respecté. Cette définition est en accord avec la définition de dissimilarité que nous avons donnée dans le chapitre 3 (section 3.1.2).

Au niveau de l'administration des médicaments anticancéreux, des différences peuvent survenir en termes de temps relatif depuis le début du cycle et/ou en termes de dosage.

Définition - dissimilarité de dose

La dissimilarité dose est définie comme :

$$\delta_d(a, a') = \frac{\text{abs}(a_d - a'_d)}{\max(a_d, a'_d)} \quad (5.6)$$

où a_d est la dose du médicament.

La notation **abs** signifie valeur absolue pour éviter toute confusion avec la notation $|S|$ qui représente la cardinalité d'un ensemble. D'après cette définition, $\delta_d(a, a')$ est un nombre réel compris entre 0 et 1. Lorsque les deux doses sont égales, la dissimilarité est nulle. Lorsque la dose administrée s'éloigne de la dose théorique, la dissimilarité augmente, et vaut 1 quand la dose n'est pas prise ou est rajoutée. Cela correspond bien à notre définition de dissimilarité.

Définition - dissimilarité du jour d'administration

La dissimilarité du jour d'administration $\delta_t(a, a')$ est défini comme suit

$$\delta_t(a, a') = \begin{cases} \text{abs}(a_t - a'_t)/\tau_{C^{theo}} & \text{si } a_t \leq \tau_{C^{theo}} \\ \text{abs}(a_t - a'_t)/\tau_{C^i} & \text{si } a_t > \tau_{C^{theo}} \\ 1 & \text{si } a = \emptyset \text{ or } a' = \emptyset \end{cases} \quad (5.7)$$

où τ_{C^i} et $\tau_{C^{theo}}$ sont respectivement les durées des cycles des administrations a et a' .

La dissimilarité du jour d'administration ne peut pas être normalisée de la même façon que la dissimilarité de dose. Elle ne peut pas être normalisée par le maximum entre le jour d'administration théorique et le jour d'administration suivi. En effet, une telle normalisation indurait une dissimilarité plus élevée lorsque les retards d'administration surviennent en début de cycle que lorsqu'ils surviennent en fin de cycle. Ainsi, dans le cas où le jour d'administration (a_t) est inférieur à la durée du cycle théorique ($\tau_{C^{theo}}$), on normalise la dissimilarité du jour d'administration avec la durée du cycle théorique. Dans le cas où le jour d'administration est supérieur à la durée du cycle théorique, on normalise par la durée du cycle suivi (τ_{C^i}). Ainsi, la dissimilarité du jour d'administration est bien comprise entre 0 et 1.

La propriété de symétrie est violée ici. La dissimilarité du jour d'administration n'est pas une distance (cf. section 3.1.2). Ceci induit que cette dissimilarité et les suivantes ne sont pas des distances.

La dissimilarité δ_{adm} entre deux administrations, ou entre une administration et l'absence d'administration (si a ou $a' = \emptyset$) est définie comme une somme pondérée des dissimilarités de temps et de dosage.

Définition - dissimilarité entre deux administrations

$$\delta_{\text{adm}}(a, a') = \begin{cases} \text{non défini} & \text{si } a_m \neq a'_m \\ \frac{\omega_t \delta_t(a, a') + \omega_d \delta_d(a, a')}{\omega_t + \omega_d} & \text{si } a_m = a'_m \\ 1 & \text{si } a = \emptyset \text{ or } a' = \emptyset \end{cases} \quad (5.8)$$

où ω_t et ω_d paramétrisent respectivement les poids associés à la dissimilarité temporelle δ_t et la dissimilarité de dose δ_d . Les poids sont définis dans l'intervalle $[0, 1]$ et ne peuvent pas être tous deux nuls.

5.2.2 Dissimilarité médicamenteux

Au niveau d'un cycle restreint au même médicament anticancéreux m , nous définissons une dissimilarité médicamenteux qui correspond à la somme des dissimilarités d'administration de ce médicament anticancéreux m . Cependant, les paramètres d'administrations d'un médicament anticancéreux m peuvent différer au sein d'un cycle : pour la même dimension a_m , il peut y avoir différents paramètres a_t et a_d . Il faut ainsi appairer administrations théoriques et administrations réelles. Par exemple, au sein d'un même cycle théorique, un médicament m peut être prévu dans trois dosages différents, à trois jours différents. Nous définissons ainsi un algorithme d'alignement pour obtenir l'ensemble aligné des paires d'administrations d'un même médicament m , en appareillant les administrations les plus proches d'un couple cycle théorique-cycle suivi.

Définition - Appariement des couples d'administration

Nous définissons \mathcal{A}_m , l'ensemble aligné des paires d'administrations théorique/réelles, en appareillant les administrations les plus proches entre le cycle théorique et le cycle suivi

$$\mathcal{A}_m = \{(a, a')_1, \dots, (a, a')_n\}.$$

Le cardinal de \mathcal{A}_m est donné par le cycle ayant le plus d'administrations soit $|\mathcal{A}_m| = \max(|\mathcal{C}^i|, |\mathcal{C}^{theo}|)$. En effet, si les cardinaux des cycles théoriques et réels sont différents, cela signifie qu'il manque une administration (la plupart du temps dans le cycle réel). Dans ce cas, l'administration concernée $(a, a')_i$ aura la forme $(a, \emptyset)_i$ ou $(\emptyset, a')_i$.

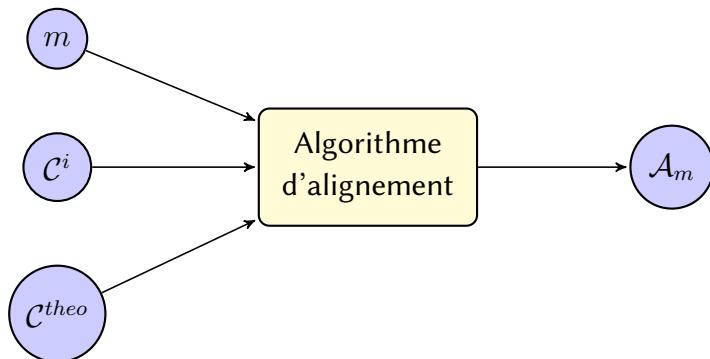


FIGURE 5.2 – Représentation scématique des entrées et sorties de l'algorithme d'alignement de ProtoDrift.

Algorithm 1 Algorithme d'alignement

```

1: Entrée :  $\mathcal{C}^i, \mathcal{C}^{theo}, m$                                 ▷ Deux cycles et un médicament
2:  $\mathcal{A}_m \leftarrow \emptyset$                                          ▷ Un alignement vide
3:  $\mathcal{C}_m^i, \mathcal{C}_m^{theo} \leftarrow \text{select}(\mathcal{C}^i, \mathcal{C}^{theo}, m)$       ▷ Réduire à un médicament particulier
4: if  $|\mathcal{C}_m^i| \leq |\mathcal{C}_m^{theo}|$  then                         ▷ Définir  $\mathcal{C}$  comme le cycle le plus court
5:    $\mathcal{C} \leftarrow \mathcal{C}_m^i$ ;  $\mathcal{C}' \leftarrow \mathcal{C}_m^{theo}$                       ▷ Définir  $\mathcal{C}'$  comme le cycle le plus long
6: else
7:    $\mathcal{C} \leftarrow \mathcal{C}_m^{theo}$ ;  $\mathcal{C}' \leftarrow \mathcal{C}_m^i$ 
8: end if
9: for  $a \in \mathcal{C}$  do                                         ▷ Pour chaque administration du cycle le plus court
10:    $\mathcal{P} \leftarrow \emptyset$ 
11:    $\delta_{min} \leftarrow 1$ 
12:   for  $a' \in \mathcal{C}'$  do
13:     if  $\delta(a, a') < \delta_{min}$  then
14:        $\mathcal{P} \leftarrow (a, a')$ 
15:        $\delta_{min} \leftarrow \delta(a, a')$ 
16:     end if
17:   end for
18:    $\mathcal{A}_m \leftarrow \mathcal{A}_m \cup \mathcal{P}$                                 ▷ Garder la paire avec la dissimilarité minimale
19: end for
20: for  $a' \in \mathcal{C}'$  do                                         ▷ Gérer les  $a'$  non appariés restants dans  $\mathcal{C}'$ 
21:   if not in( $a', \mathcal{A}_m$ ) then
22:      $\mathcal{A}_m \leftarrow \mathcal{A}_m \cup (\emptyset, a')$ 
23:   end if
24: end for
25: Sortie :  $\mathcal{A}_m$ 
  
```

Définition - dissimilarité médicament

Nous définissons la dissimilarité entre le cycle théorique et réel pour un médicament m (un couple m (molécule, mode)) comme la moyenne des dissimilarités des adminis-

trations de ce médicament

$$\delta_m(\mathcal{C}^i, \mathcal{C}^{theo}) = \sum_{(a,a') \in \mathcal{A}_m} \frac{\delta(a, a')}{\max(|\mathcal{C}_m^i|, |\mathcal{C}_m^{theo}|)}. \quad (5.9)$$

Afin de calculer la dissimilarité totale associée à un cycle, on somme sur les dissimilarités des différents médicaments qui composent ce cycle. Au niveau d'un cycle, il peut y avoir une dissimilarité de médicament δ_m pour chaque médicament m (molécule, mode).

5.2.3 Dissimilitarité intra-cycle et dissimilitarité inter-cycle

Définition - dissimilitarité intra-cycle

La dissimilitarité intra-cycle δ_{intra} s'écrit

$$\delta_{\text{intra}}(\mathcal{C}^i, \mathcal{C}^{theo}) = \frac{\sum_{m \in \mathcal{M}} \omega_m \delta_m(\mathcal{C}^i, \mathcal{C}^{theo})}{\sum_{m \in \mathcal{M}} \omega_m} \quad (5.10)$$

où les coefficients $\omega_m \in [0, 1]$ sont les poids associés à chaque médicament. Ils ne peuvent pas être tous deux nuls afin que le dénominateur soit strictement positif.

La dissimilitarité intra-cycle est bien définie entre 0 et 1. Comme nous le verrons ensuite, dans le cadre de cette thèse, le poids devant chaque molécule sera égal : on accorde la même importance à chaque molécule. Dans ce cas, la dissimilitarité intra-cycle s'écrit

$$\delta_{\text{intra}}(\mathcal{C}^i, \mathcal{C}^{theo}) = \sum_{m \in \mathcal{M}} \frac{1}{|\mathcal{M}|} \delta_m(\mathcal{C}^i, \mathcal{C}^{theo}). \quad (5.11)$$

Nous définissons maintenant la dissimilitarité inter-cycle, qui prend en compte l'inter-cure, le décalage temporel entre deux cycles : jusque là, la dimension temporelle des administrations était relative au début du cycle. Nous voudrions maintenant prendre en compte la dissimilitarité associée au décalage temporel entre deux cycles : il s'agit là de prendre en compte la durée d'inter-cure que nous avons définie dans la section 1.1.4 du chapitre 1.

Prenons un exemple : après quatre cycles de chimiothérapie en décembre, il est décidé de repousser le début du cinquième cycle après les vacances de Noël pour que le patient puisse passer Noël en famille. Cela signifie que le cinquième cycle va démarrer après sa date de début théorique. Une possibilité est donc de définir la dissimilitarité inter-cycle comme la différence entre la date de début du cycle i et la date théorique du cycle associé. Cette définition par date pose un problème : un retard au cycle 5 implique un retard pour les cycles suivants (6, 7, etc...). Nous voudrions donc que la dissimilitarité associée au retard du cycle i ne dépende pas des retards associés aux cycles précédents. Le retard au cycle i est ainsi dû à l'augmentation de la durée d'inter-cure qui fait partie du cycle $i - 1$.

Définition - dissimilitarité inter-cycle

La dissimilarité entre deux cycles est définie comme

$$\delta_{\text{inter}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}}) = \begin{cases} 0 & \text{si } i = 1 \\ \frac{\text{abs}(\tau_{\mathcal{C}^{i-1}} - \tau_{\mathcal{C}^{\text{theo}}})}{\max(\tau_{\mathcal{C}^{i-1}}, \tau_{\mathcal{C}^{\text{theo}}})} & \text{si } i > 1 \end{cases} \quad (5.12)$$

où $\tau_{\mathcal{C}^{i-1}}$ et $\tau_{\mathcal{C}^{\text{theo}}}$ sont respectivement les durées du cycle $i - 1$ et du cycle théorique.

La dissimilarité inter-cycle est donc bien définie entre 0 et 1. Elle reflète le retard du cycle i dû à l'inter-cure du cycle $i - 1$ plus longue que l'inter-cure du cycle théorique. De cette façon, le retard du cycle i ne se propage pas aux cycles $j > i$.

5.2.4 Dissimilarités : du cycle à la ligne

Nous pouvons donc maintenant définir la dissimilarité totale associée au cycle i . Il s'agit de la moyenne pondérée de la dissimilarité inter-cycle et intra-cycle.

Définition - dissimilarité cycle

La dissimilarité cycle δ_{cycle}

$$\delta_{\text{cycle}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}}) = \frac{\omega_{\text{intra}} \delta_{\text{intra}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}}) + \omega_{\text{inter}} \delta_{\text{inter}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}})}{\omega_{\text{intra}} + \omega_{\text{inter}}} \quad (5.13)$$

où ω_{intra} et ω_{inter} sont respectivement les poids associés aux dissimilarités intra-cycle et inter-cycle. Les poids sont définis dans l'intervalle $[0, 1]$ et ne peuvent pas être tous deux nuls.

Nous sommes ces dissimilarités afin de définir la dissimilarité cumulative au cycle k .

Définition - dissimilarité cumulative au cycle k

La dissimilarité cumulative au cycle k δ_{cycle} est définie comme

$$\delta_{\text{acc}}(\mathcal{C}^k, \mathcal{C}^{\text{theo}}) = \frac{1}{k} \sum_{i=1}^k \delta_{\text{cycle}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}}) \quad (5.14)$$

Enfin la dissimilarité de la ligne δ_{line} est définie comme la dissimilarité cumulative au dernier cycle.

Définition - dissimilarité de ligne

La dissimilarité entre une ligne \mathcal{L} et un protocole prévu (*i.e.*, une séquence de cycles théoriques)

$$\delta_{\text{line}}(\mathcal{L}, \mathcal{C}^{\text{theo}}) = \frac{1}{l} \sum_{i=1}^l \delta_{\text{cycle}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}}) = \delta_{\text{acc}}(\mathcal{C}^l, \mathcal{C}^{\text{theo}}) \quad (5.15)$$

où l est le nombre total de cycles de la ligne \mathcal{L} (ou le dernier numéro de cycle de la ligne).

La dissimilarité de ligne δ_{line} est bien comprise entre 0 et 1. Elle donne une valeur agrégée de l'ensemble des dissimilarités, c'est-à-dire des écarts au protocole théorique survenu lors de la ligne de traitement suivie.

5.2.5 Exemple de calcul de dissimilarité

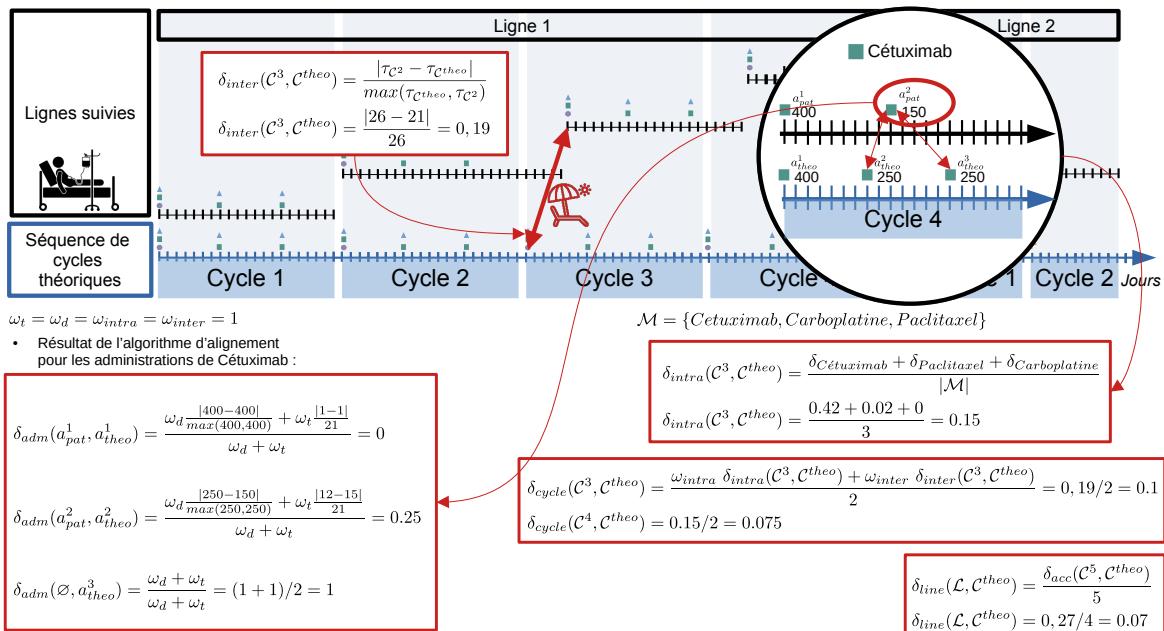


FIGURE 5.3 – Calcul des différentes dissimilarités mesurant la dérive de première ligne suivie par le patient fictif de la figure 5.1.1.

Un exemple de calcul de dissimilarité est donné en figure 5.3. On reprend le scénario d'adhésion exposée dans la figure 5.1.1, et les dissimilarités calculées lors de la première ligne de traitement du patient fictif. La ligne suivie est représentée par les séquences en escalier noir, où chaque palier représente un cycle suivi. Le protocole de chimiothérapie, c'est-à-dire la séquence de cycles théoriques décrits par le schéma thérapeutique 1465, est représentée en bleue.

On fixe tous les poids associés aux dissimilarités à 1.

Une première dissimilarité inter-cycle est calculée au cycle 3, qui est retardé par rapport au cycle 3 de la séquence de cycle théorique. Le cycle suivi 2 a duré plus longtemps que le cycle théorique (respectivement 26 jours et 21 jours), avec une plus longue période d'inter-cure. La dissimilarité inter du cycle 3 est de 0.19. Les administrations du cycle 3, sont identiques au cycle théorique, donc la dissimilarité intra-cycle du cycle 3 est nulle. Les poids intra-cycle et inter-cycle étant égaux comme tous les autres poids à 1, la dissimilarité cycle est la moyenne d'une dissimilarité inter-cycle de 0.19 et d'une dissimilarité intra-cycle nulle. Elle vaut donc $0.19/2 = 0.1$.

Ensuite, au cycle 4 le patient reçoit deux administrations de Cétuximab à la place d'en recevoir trois, comme c'est le cas dans le cycle théorique. Les dissimilarités d'administration sont calculées pour chaque paire d'administration théorique-suivie possible. Chacune des dissimilarités d'administration est composée d'une dissimilarité de dose et de jour d'administration. L'algorithme d'alignement retourne les trois paires d'administration ayant une dissimilarité d'administration minimale, l'ensemble $((a_{pat}^1, a_{theo}^1), (a_{pat}^2, a_{theo}^2), (\emptyset, a_{theo}^3))$. Il s'agit :

- de la première administration réelle et la première administration théorique. Ici, La dose et le jour correspondent : la dissimilarité est nulle ;
- de la deuxième administration réelle et la deuxième administration théorique. Dans ce cas, la dissimilarité administration a une contribution temporelle (administration réelle le dixième jour alors qu'elle est prévue le huitième jour) et une dissimilarité de dose (150 contre 250). La dissimilarité administration vaut alors 0.25 ;
- de la troisième administration théorique qui n'a pas été effectuée. Dans ce cas, la dissimilarité vaut 1.

La dissimilarité Cétuximab associée à ce cycle vaut donc $1.25/3 = 0.42$. La dissimilarité associée au Paclitaxel administré au jour 10, au lieu du jour 8 est de 0.02. La dissimilarité associée au Carboplatine est nulle. La dissimilarité intra-cycle est donc égale à $0.44/3$ (trois médicaments différents).

Au cours de cette ligne, seule les cycles 3 et 4 ont une dissimilarité non nulle. Cette première ligne est composée de quatre cycles donc la dissimilarité ligne vaut $0.27/3 = 0.007$.

La méthodologie de ProtoDrift peut être visualisée comme un arbre hiérarchique, illustrant la relation entre les différents niveaux de dissimilarités. Cette structure arborescente capture l'essence des décisions des médecins à différents stades et périodes de traitement, fournissant une vue d'ensemble des écarts de traitement.

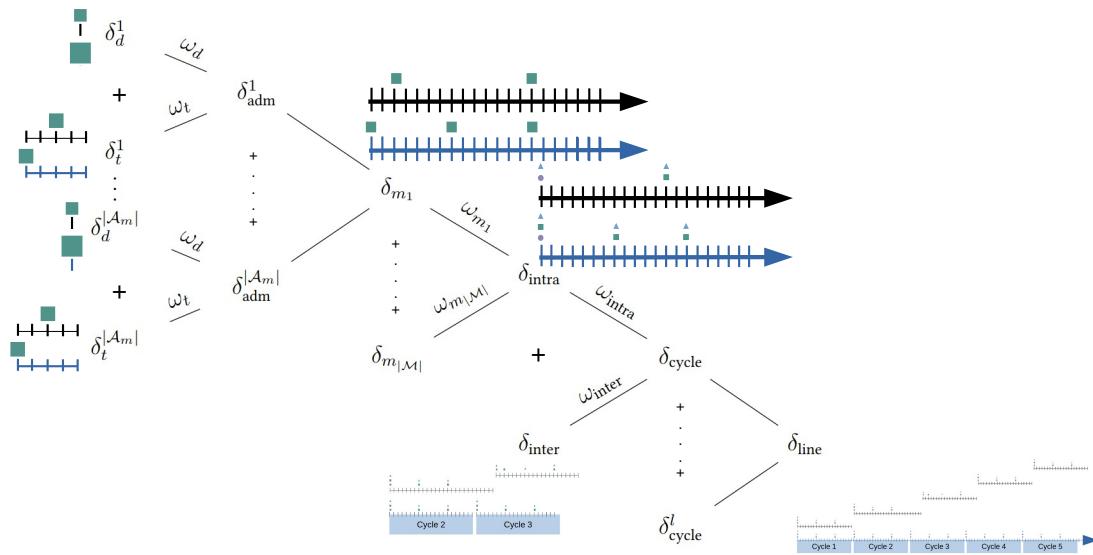


FIGURE 5.4 – Représentation schématique, sous forme d’arbre horizontal, des dépendances entre les différentes dissimilarités qui composent ProtoDrift. Sur la gauche, les feuilles de l’arbre correspondent aux dissimilarités temporelles et de dose d’administration. Les dissimilarités sont sommées à chaque nœud, correspondant à différentes étapes du traitement, jusqu’à la dissimilarité ligne, qui est la racine de l’arbre, sur la droite du schéma. Chaque dissimilarité au niveau des nœuds est la somme normalisée de ses dissimilarités enfants. $|\mathcal{A}_m|$ correspond au maximum entre les administrations de cycles réels et théoriques d’un médicament anticancéreux m . $|\mathcal{M}|$ correspond au nombre de médicaments anticancéreux distincts dans le cycle. l correspond au nombre de cycles réels suivis par les patients dans la ligne de traitement.

Notation - ProtoDrift Naïf (NP)

Dans la suite du chapitre, on nomme "ProtoDrift Naif" (NP pour *Naive ProtoDrift*) la méthode qui consiste à utiliser ProtoDrift avec des poids associés à chacune des dissimilarités tous égaux à 1 ($\omega_t = \omega_d = \omega_m = \omega_{intra} = \omega_{inter} = 1$) pour mesurer l’adhésion aux chimiothérapies.

Dans la section suivante, nous nous intéressons justement à optimiser les valeurs de ces poids, et étudier leur impact sur la survie globale.

5.3 Optimisation des poids de ProtoDrift

5.3.1 Introduction à l’optimisation

L’optimisation des poids de ProtoDrift vise à calibrer ces poids pour améliorer la pré-diction de la survie globale des patients par rapport aux modèles standards. Ce processus

repose sur l'exploration systématique des différentes combinaisons de poids en utilisant une stratégie de recherche par grille (*grid search*). L'objectif principal est de trouver le meilleur équilibre entre les dissimilarités temporelles et de dosage.

Pour assurer la robustesse des résultats, nous utilisons un ré-échantillonnage *bootstrap*, ce qui permet de prendre en compte la variabilité au sein de la population de patients. Les performances associées à chaque combinaison de valeur testée pour les poids sont évaluées à l'aide de mesures statistiques adaptées au modèle de régression utilisé : soit la régression logistique, soit le modèle de Cox. À chaque étape de la recherche par grille, nous évaluons la performance prédictive de ProtoDrift, ce qui nous permet d'identifier la combinaison optimale de poids offrant la meilleure performance dans les prédictions de survie. Cette optimisation permet non seulement d'améliorer les scores de prédiction, mais aussi d'explorer les combinaisons de poids les plus influentes sur la survie globale.

5.3.2 Objectifs de l'optimisation

ProtoDrift intègre divers poids associés aux dissimilarités définies à différents niveaux. Ceux-ci reflètent l'importance relative des modifications apportées à un protocole de chimiothérapie. Dans cette étude, nous nous concentrons particulièrement sur les poids qui reflètent les décisions médicales au sein d'une ligne de traitement, et qui visent à améliorer le confort du patient, sa qualité et/ou ses chances de survie. En particulier, nous nous intéressons aux ajustements des doses des médicaments anticancéreux, à leur suppression et aux décalages temporels, que ce soit au sein d'un cycle ou entre les cycles.

Nous revenons ici sur la définition de la dissimilarité intra-cycle (*cf. formule 5.11*) et de la prise en compte du poids des médicaments ω_m . La mesure ProtoDrift et l'algorithme implémenté laisse la possibilité de définir des pondérations de médicaments. Dans le cas où la taille des cohortes de patients est grande, il est possible de coupler ProtoDrift à des modèles d'apprentissage statistique, afin d'optimiser la survie sur les poids des médicament. Changer la dose d'un médicament plutôt qu'un autre a très probablement une influence différente sur les réponses au traitement. De même décaler le jour d'administration d'un médicament plutôt qu'un autre, a certainement des conséquences différentes. Notons, qu'il est peu commun en apprentissage automatique, de s'intéresser plus aux poids qu'aux performances de prédiction. Dans notre cas, l'importance relative des poids médicament aurait un sens médical : décaler un médicament plutôt qu'un autre affecte-t-il la survie du patient ? Ce type de déviation au protocole théorique et son impact sur la survie seraient intéressants à étudier.

Dans l'application de l'optimisation aux données de l'HEGP et du CHU de Bordeaux (section 5.7), nous verrons que nous sélectionnons les localisations de cancer pour lesquelles il y a au moins 400 patients traités. Pour une localisation de cancer, il y a une dizaine de médicaments différents. La cohorte de patients doit ensuite être répartie en un ensemble d'entraînement et de validation. Ce nombre de patients est trop faible pour introduire autant de paramètres libres.

Remarque - poids molécule

Dans la suite de ce travail, nous prenons un poids de médicaments ω_m de 1, que nous n'optimisons pas.

Les leviers de décision médicale que nous mesurons sont les suivants.

- Les ajustements au sein d'un cycle :
 - Les variations de dose de l'administration de la molécule anticancéreuse.
 - Les variations temporelles de l'administration de la molécule anticancéreuse.
- Les retards de cycle.

Dans ProtoDrift, ces ajustements sont mesurés par les dissimilarités δ_d , δ_t , δ_{intra} et δ_{inter} , et leur importance relative peut être ajustée grâce aux poids associés : ω_d , ω_t , ω_{intra} et ω_{inter} .

L'objectif de l'optimisation se décline en trois points :

- Identifier les combinaisons de poids associés à de meilleures performances que celles des méthodes NP et BADRDI.
- Explorer les combinaisons de poids associant le temps et le dosage et leur corrélation avec les résultats de survie.
- Adapter ProtoDrift à différentes localisations de cancer et étudier les différences de ces deux premiers objectifs selon les localisations.

Notez que notre objectif principal est d'évaluer si ProtoDrift peut mieux prédire la survie que la méthode BADRDI, plutôt que de prédire précisément la survie dans l'absolu.

Les dissimilarités de ProtoDrift sont des mesures longitudinales et peuvent être prises à différents niveaux du traitement de chimiothérapie. Nous évaluons ProtoDrift avec la dissimilarité ligne-protocole (δ_{line}). Premièrement, parce que, comme nous l'avons mentionné en début de section, nous voulons prendre en compte les ajustements se produisant au sein d'une ligne de traitement. Deuxièmement, parce que les études effectuées avec la RDI calculent généralement sa valeur sur une ligne de traitement.

Enfin, l'optimisation est effectuée sur les prédictions de survie globale. De nombreux autres résultats de survie peuvent être utilisés, tels que la survie sans progression ou la survenue de toxicités irréversibles. La survie globale est un résultat que nous pouvons obtenir de manière quasi-systématique avec une haute qualité 1.4.8. De plus, nous voulons imiter les études réalisées avec la RDI, qui utilisent généralement la survie globale pour étudier les impacts de la non-adhésion.

5.3.3 Méthode d'optimisation

5.3.3.1 Stratégie globale

Nous souhaitons déterminer la combinaison de poids (ω_d , ω_t , ω_{intra} , ω_{inter}) qui conduit aux meilleurs gains de performance dans la prédiction de survie par rapport aux performances obtenues avec la méthode BADRDI. Pour ce faire, nous explorons différentes combinaisons de poids à l'aide d'une recherche par grille (*grid search*), calculons la dissimilarité ProtoDrift résultante δ_{line} , ajustons et prédisons des modèles de régression logistique et de Cox, puis comparons les scores de prédiction obtenus.

5.3.3.2 Définition de la grille

Les dissimilarités que nous calculons via les paramètres ($\omega_d, \omega_t, \omega_{\text{intra}}, \omega_{\text{inter}}$) ne sont pas indépendantes. Par exemple, la dissimilarité de ligne calculée dans le cas où tous les paramètres seront égaux à 1 est la même que celle où tous les paramètres sont égaux à 2. La dissimilarité finale dépend de la différence relative entre les poids et non de leur valeur absolue. D'un espace à quatre paramètres, nous pouvons donc réduire notre étude à un espace à deux paramètres.

Définition - α

Nous définissons ainsi le paramètre α qui correspond au poids relatif entre les décalages en jours et en dose des administrations au sein des cycles :

$$\alpha = \frac{\omega_t}{(\omega_t + \omega_d)} \in [0, 1]. \quad (5.16)$$

Dans le cas où α vaut 1, cela signifie que les dissimilarités de dose ne sont pas prises en compte. Dans le cas où il est nul, ce sont les dissimilarités jours qui ne sont pas prises en compte.

Définition - β

Le paramètre β correspond au poids relatif entre les dissimilarités inter et intra-cycles :

$$\beta = \frac{\omega_{\text{inter}}}{(\omega_{\text{inter}} + \omega_{\text{intra}})} \in [0, 1] \quad (5.17)$$

Lorsque β est nul, seules les dissimilarités intra-cycle participent à la dissimilarité totale. Plus β est grand, plus les retards inter-cycles contribuent à la dissimilarité.

Propriété - Dissimilarités en fonction de α et β

Les dissimilarités δ_{adm} et δ_{cycle} peuvent être réécrites en fonction de $\alpha, \beta \in [0, 1]$:

$$\delta_{\text{adm}}(a, a') = \alpha \delta_t(a, a') + (1 - \alpha) \delta_d(a, a') \quad (5.18)$$

$$\delta_{\text{cycle}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}}) = \beta \delta_{\text{inter}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}}) + (1 - \beta) \delta_{\text{intra}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}}) \quad (5.19)$$

Démonstration. Étant donnée la définition de α , nous pouvons exprimer ω_d en fonction de ω_t et α :

$$\omega_d = \frac{\omega_t(1 - \alpha)}{\alpha}$$

ce qui permet d'écrire la dissimilarité administration

$$\begin{aligned}\delta_{\text{adm}}(a, a') &= \left(\omega_t \cdot \delta_t(a, a') + \frac{\omega_t(1 - \alpha)\delta_d(a, a')}{\alpha} \right) \frac{1}{\omega_t + \omega_d} \\ &= \alpha\delta_t(a, a') + \frac{\omega_t}{\alpha(\omega_t + \omega_d)}(1 - \alpha)\delta_d(a, a') \\ &= \alpha\delta_t(a, a') + (1 - \alpha)\delta_d(a, a')\end{aligned}$$

qui ne dépend que de α . Le même raisonnement peut être appliqué pour l'expression de la dissimilarité cycle avec β . \square

Afin d'explorer l'ensemble des paramètres, nous parcourons l'espace à deux dimensions $\{\alpha, \beta\} = [0, 1] \times [0, 1]$. En pratique, nous parcourons cette grille avec un certain pas et calculons les dissimilarités ProtoDrift résultantes. Nous ajustons et prédisons alors deux modèles.

5.3.3.3 Modèles de régression

– **Modèle de régression logistique :**

$$\text{logit}(P(Y = 1)) = \theta_0 + \theta_1 \text{metric} + \theta_2 \text{age} + \theta_3 \text{sex}$$

– **Modèle de régression de Cox :**

$$h(t) = h_0 \exp(\theta_1 \text{metric} + \theta_2 \text{age} + \theta_3 \text{sex})$$

Dans les deux formules, la variable explicative metric correspond, pour un numéro de ligne donné, soit à la dissimilarité ProtoDrift ligne-protocole δ_{line} , soit à l'ADRDI.

Dans la régression logistique, la variable de réponse binaire Y encode la survenue de l'événement avant la fin de l'étude, tandis que dans le modèle de régression de Cox, $h(t)$ représente la fonction de hasard à l'instant t (*cf. section 3*)

Pour les localisations de cancer génitales, le dernier terme est absent des deux formules ($\theta_3 \times \text{sex}$).

5.3.3.4 Rééchantillonnage Bootstrap

Nous introduisons le rééchantillonnage *bootstrap* avec remise dans notre processus pour prendre en compte la variabilité des populations échantillonées de patients et évaluer l'incertitude des prédictions. L'objectif de ce rééchantillonnage *bootstrap* est de tenir compte des effets spécifiques à la population. Le rééchantillonnage est effectué sur l'ensemble du jeu de données avant de le diviser en ensembles d'entraînement et de test. Comme nous voulons étudier l'effet de différentes combinaisons de poids, les mêmes échantillons *bootstrap* doivent être utilisés à chaque étape de la recherche par grille.

5.3.3.5 Paramètres personnalisables de l'algorithme d'optimisation

Plusieurs types de paramètres sont impliqués dans l'algorithme d'optimisation.

– **Paramètres d'adaptation** : sélection d'une population spécifique sur laquelle nous voulons adapter ProtoDrift.

– **Localisation du cancer** : la localisation de la tumeur.

- **Numéro de ligne** : le numéro de ligne auquel la dissimilarité ProtoDrift ligne-protocole δ_{line} ou l'intensité de dose relative de tous les médicaments ADRDI sont calculées pour ajuster les régressions.
- **Paramètre de résultat de l'optimisation** : détermine ce que nous prédisons.
- **Temps de survie** : le temps relatif entre la date de début de la ligne et la fin de l'étude. Par exemple 3 ans.
- **Paramètres de l'algorithme** :
 - **Étape de recherche par grille** : le pas pour parcourir les valeurs de α et β
 - **Nombre d'échantillons bootstrap** : le nombre de fois où la population de patients est divisée en ensembles d'entraînement et de validation.
 - **Pourcentage de répartition** : la proportion des patients divisée entre ensembles d'entraînement et de validation. Par exemple 70% d'entraînement et 30% de validation.
- **Paramètre de stratification** :
 - **Tranches d'âge** : définition des catégories d'âge pour la variable de stratification. Par exemple des tranches d'âge de 5ans.

5.3.3.6 Évaluation des prédictions

Les prédictions sont évaluées à l'aide de métriques couramment utilisés pour les deux modèles de régression.

- **Scores de prédition de la régression logistique** :
 - Aire sous la courbe de ROC (AUC-ROC) : varie entre 0 et 1. Un score de 0,5 indique une prédition aléatoire, tandis qu'un score de 1 indique une parfaite discrimination entre les classes.
 - Coefficient de corrélation de Matthews (MCC) : varie entre -1 et 1. Un score de 1 indique une parfaite prédition, 0 indique une prédition aléatoire, et -1 indique une prédition totalement incorrecte.
 - F1-score : varie entre 0 et 1. Un score de 0 indique une performance très faible, tandis qu'un score de 1 indique une performance parfaite en termes de précision et de rappel. C'est la moyenne harmonique de la précision et du rappel.
- **Scores de prédition du modèle de régression de Cox** :
 - Indice de concordance (C-index) : varie entre 0,5 et 1. Un score de 0,5 indique une performance aléatoire, tandis qu'un score de 1 indique une parfaite concordance entre les prédictions et les observations.
 - Score de Brier : varie entre 0 et 1. Un score de 0 indique une prédition parfaite, tandis qu'un score de 1 indique une prédition complètement incorrecte.

Remarque - scores de prédition

Si tous les scores ont été calculés, dans la suite de ce travail, on présente seulement les résultats obtenus avec l'AUC-ROC pour la régression logistique et le C-index pour la

régression de Cox.

5.3.3.7 Stratégie détaillée

Filtrer la table ProtoDrift selon la localisation du cancer et le numéro de ligne

Nous filtrons la table de départ ProtoDrift (*cf.* figure 5.5) selon la localisation du cancer et le numéro de la ligne de traitement pour obtenir une table avec une entrée par patient. Ensuite, cette table est jointe aux données de survie. La table de données résultante est composé d'une entrée par patient avec la valeur δ_{line} et les données de survie (*cf.* figure 5.6).

PAT_NUM	AGE	SEX	LOCATION	LINE_NUM	CYCLE_NUM	molecule	adm_mode	adm_id	$\delta_t(a, a')$	$\delta_i(a, a')$	$\delta_m(a, a')$	$\delta_m(C^t, C^{t\text{bleu}})$	$\delta_{\text{intra}}(C^t, C^{t\text{bleu}})$	$\delta_{\text{inter}}(C^t, C^{t\text{bleu}})$	$\delta_{\text{acc}}(C^t, C^{t\text{bleu}})$	$\delta_{\text{line}}(C^t, C^{t\text{bleu}})$	line_all_drugs_RDI		
P1	48	F	colon	L1	C1	FLUOROURACILE	bolus	A1	0	0.00297	0.00149	0.00149	0.66716	0.0	0.33358	0.33358	0.45358	0.84545	
					C1	FLUOROURACILE	low	A2	1.0	1.0	1.0	1.0	0.66716	0.0	0.33358	0.33358	0.45358	0.84545	
					C1	OXALIPLATINE	low	A3	1.0	1.0	1.0	1.0	0.66716	0.0	0.33358	0.33358	0.45358	0.84545	
					C2	FLUOROURACILE	bolus	A1	0	0.00297	0.00149	0.00149	0.66716	0.0	0.33358	0.66716	0.45358	0.84545	
					C2	FLUOROURACILE	low	A2	1.0	1.0	1.0	1.0	0.66716	0.0	0.33358	0.66716	0.45358	0.84545	
					C2	OXALIPLATINE	low	A3	1.0	1.0	1.0	1.0	0.66716	0.0	0.33358	0.66716	0.45358	0.84545	
						
						
						
					L1	CB	FLUOROURACILE	bolus	A1	0	0.0	0.0	0.0	
					(Last)		
					L1	CB	FLUOROURACILE	low	A2	0	0.0	0.0	0.0	
					(Last)		
					L1	CB	OXALIPLATINE	low	A3	0	0.00355	0.00178	0.00178	0.00059	0.0	0.000295	3.62864	0.45358	0.84545
					L2	C1	FLUOROURACILE	bolus	A1	0	0.00297	0.00149	0.00149	0.66716	0.0	0.33358	0.33358	0.23568	0.96882
						
						
						
			breast	L1	C1	PACLITAXEL	low	A1	0	0.00485	0.00286	0.002425	0.58948	0.28928	0.43958	0.43958	0.0398	0.9666	
				
				
				
				
P2	64	M	colon	L1	C1	FLUOROURACILE	bolus	A1	0	0.00297	0.00297	0.00149	0.66716	0.0	0.33358	0.33358	0.23558	0.1253	
.		
.		
.		

FIGURE 5.5 – Structure de la table ProtoDrift de départ, avec une entrée par administration. Tous les poids associés aux dissimilarités sont égaux à 1. Les dissimilarités de cette table correspondent aux dissimilarités du "ProtoDrift Naïf" (NP) ($\omega_d = \omega_t = \omega_{\text{intra}} = \omega_{\text{inter}} = 1 \Leftrightarrow \alpha = \beta = 1/2$). Les "|" signifient une valeur identique à l'entrée supérieure. Les ":" signifient que les valeurs peuvent changer. La table est groupée par patient, localisation, ligne de traitement, cycle et médicament. Les dissimilarités sont colorées en fonction du niveau d'arborescence auquel elles sont associées (*cf.* figure 5.4) : jaune foncé pour la ligne, bleu pour le cycle, vert pour le sous-ensemble de cycle restreint aux administrations d'un même médicament, rose pour l'administration. La couleur verte a été choisie pour le niveau de δ_m car les médicaments, paire (molécule, mode d'administration), sont colorés en jaune clair, et les cycles en bleu.

PAT_NUM	AGE	SEX	$\delta_{line}(\mathcal{L}, \mathcal{C}^{theo})$	line_all_drugs_RDI	Time	status
P1	48	F	0.45358	0.84545	203	2
P2	64	M	0.25358	0.12533	1095	1
P3	36	M	0.51264	0.63971	1095	1
.
.
.
PN	64	M	0.36541	0.98236	928	2

FIGURE 5.6 – La table ProtoDrift Naïf (figure 5.5) est filtrée selon une localisation de cancer et un numéro de ligne de traitement, et associée aux données de survie. La table contient donc une entrée par patient.

Définition des échantillons *bootstrap*: À partir de ce jeu de données initial, nous créons plusieurs échantillons *bootstrap*. Chaque échantillon *bootstrap* est ensuite divisé en ensembles d'entraînement et de test. Ces paires d'ensembles d'entraînement et de test sont stockées et utilisées pour ajuster et prédire les modèles de régression.

Initialisation Pour chaque échantillon *bootstrap* :

- Nous ajustons des modèles de régression logistique et de Cox avec l'ADRDI et δ_{line} en variable explicative (*cf.* section 5.3.3.3) sur l'ensemble d'entraînement et stockons les estimations des paramètres.
- Nous évaluons ces modèles en calculant les scores de prédiction sur l'ensemble de test et les stockons.

Nous calculons la moyenne, l'écart-type et les intervalles de confiance des estimations des modèles et des scores de prédiction.

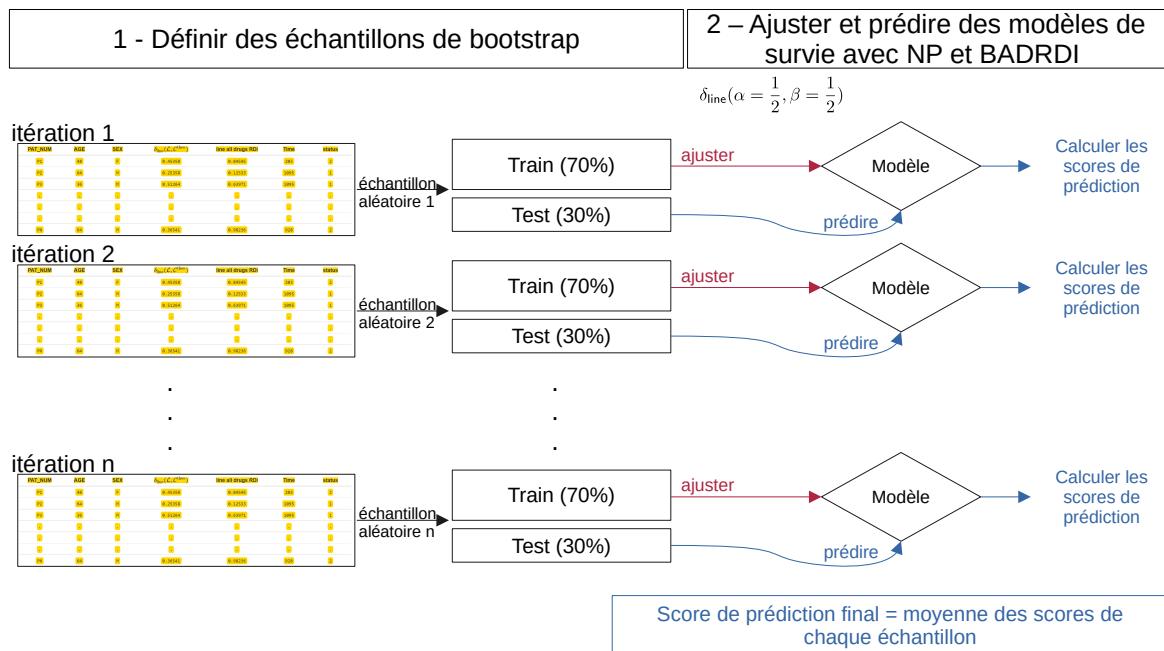


FIGURE 5.7 – Étapes 1 et 2 de l’algorithme d’optimisation. Au cours de la première étape, des échantillons *bootstrap* sont définis à partir des numéros de patients de la table Proto-Drift Naïf filtrée (*cf.* figure 5.6). Au cours de la deuxième étape, des modèles de régression logistique et de Cox avec $\delta_{\text{line}}(\alpha = \frac{1}{2}, \beta = \frac{1}{2})$ et ADRDI comme variables explicatives (*cf.* section 5.3.3.3) sont ajustés et prédits, et leurs performances de prédiction sont évaluées. La performance finale des modèles correspond à la moyenne des scores de prédiction obtenus sur les échantillons *bootstrap*. La réalisation de ces deux étapes permettent de réaliser la double analyse comparative NP-BADRDI que nous verrons dans la section 5.4.

Boucle sur les combinaisons α - β : Pour chaque combinaison α, β :

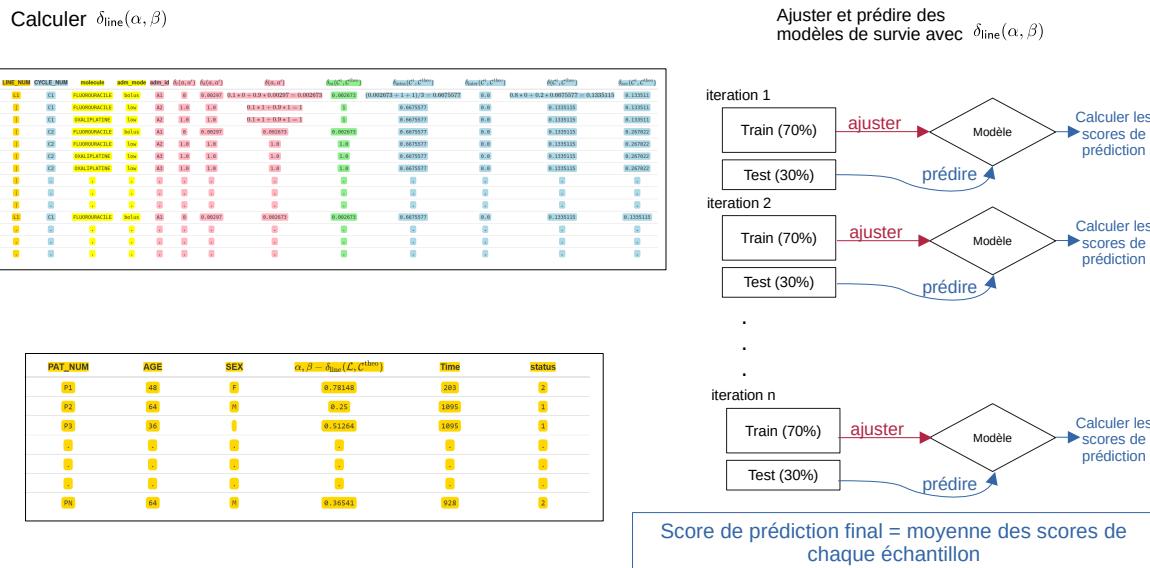
Nous calculons la dissimilarité ProtoDrift ligne-protocole $\delta_{\text{line}}(\alpha, \beta)$ résultante .

Pour chaque échantillon *bootstrap* :

- Nous ajustons des modèles de régression logistique et de Cox avec $\delta_{\text{line}}(\alpha, \beta)$ sur l’ensemble d’entraînement et stockons les estimations des paramètres.
- Nous évaluons ces modèles en calculant les scores de prédiction sur l’ensemble de test et stockons ces scores.

Nous calculons la moyenne, l’écart-type et les intervalles de confiance des estimations des modèles et des scores de prédiction pour chaque combinaison α, β .

3 – Tourner sur les valeurs α, β



Notation - ProtoDrift Optimisé

PS-OP : AUC-ROC-OP ou C-index-OP

Dans la suite du chapitre, on nommera *Prediction Score Optimised ProtoDrift* (PS-OP) la méthode qui consiste à utiliser la variable $\operatorname{argmax}_{\alpha, \beta} \text{PS}(\delta_{\text{line}}(\alpha, \beta))$, pour mesurer l'adhésion aux chimiothérapies.

- Lorsque le score d'optimisation est l'AUC-ROC, on notera AUC-ROC-OP.
- Lorsque le score d'optimisation est le C-index, on notera C-index-OP.

La méthode PS-OP est le résultat direct de l'optimisation. Nous sélectionnons les paires (α, β) de la grille de recherche qui maximisent un score de prédiction. Pour la régression logistique, nous sélectionnons les paires (α, β) qui maximisent l'aire sous la courbe de ROC (AUC-ROC), et pour la régression de Cox, nous sélectionnons les paires (α, β) qui maximisent l'indice de concordance (C-index). Les modalités des trois modèles sont résumées dans les tableaux 5.1 et 5.2.

Nom de la méthode évaluée	Régression logistique	
	Notation de la variable explicative	Formulation du modèle
Baseline All Drugs Relative Dose-Intensity (BADRDI)	ADRDI	$\text{logit}(P(Y_i = 1)) = \theta_0 + \theta_1 \text{ADRDI} + \theta_2 \text{age} + \theta_3 \text{sex}$
ProtoDrift Naïf (NP)	$\delta_{\text{line}}(\alpha = \frac{1}{2}, \beta = \frac{1}{2})$	$\text{logit}(P(Y_i = 1)) = \theta_0 + \theta_1 \delta_{\text{line}}(\alpha = \frac{1}{2}, \beta = \frac{1}{2}) + \theta_2 \text{age} + \theta_3 \text{sex}$
ProtoDrift optimisé par AUC-ROC (AUC-ROC-OP)	$\operatorname{argmax}_{\alpha, \beta} \text{AUC}(\delta_{\text{line}}(\alpha, \beta))$	$\text{logit}(P(Y_i = 1)) = \theta_0 + \theta_1 \delta_{\text{line}}(\operatorname{argmax}_{\alpha, \beta} \text{AUC}) + \theta_2 \text{age} + \theta_3 \text{sex}$

TABLE 5.1 – Modèles de régression logistique pour BADRDI, NP et AUC-ROC-OP

Nom de la méthode évaluée	Régression de Cox	
	Notation de la variable explicative	Formulation du modèle
Baseline All Drugs Relative Dose-Intensity (BADRDI)	ADRDI	$h(t) = h_0(t) \times \exp(\theta_1 \text{ADRDI} + \theta_2 \text{age} + \theta_3 \text{sex})$
ProtoDrift naïf (NP)	$\delta_{\text{line}}(\alpha = \frac{1}{2}, \beta = \frac{1}{2})$	$h(t) = h_0(t) \times \exp(\theta_1 \delta_{\text{line}}(\alpha = \frac{1}{2}, \beta = \frac{1}{2}) + \theta_2 \text{age} + \theta_3 \text{sex})$
ProtoDrift optimisé par C-index (C-index-OP)	$\operatorname{argmax}_{\alpha, \beta} \text{C-index}(\delta_{\text{line}}(\alpha, \beta))$	$h(t) = h_0(t) \times \exp(\theta_1 \delta_{\text{line}}(\operatorname{argmax}_{\alpha, \beta} \text{C-index}) + \theta_2 \text{age} + \theta_3 \text{sex})$

TABLE 5.2 – Modèles de régression de Cox pour BADRDI, NP et C-index-OP

Nous évaluons la performance des trois modèles -Baseline All Drugs RDI (BADRDI), ProtoDrift naïf (NP) et ProtoDrift optimisé par score de prédiction (PS-OP) - selon trois

dimensions principales :

- **Significativité de la contribution de la variable explicative au modèle de régression :** Nous comparons la significativité des contributions de l'ADRDI, de $\delta_{\text{line}}(\alpha = \frac{1}{2}, \beta = \frac{1}{2})$ et de $\text{argmax}_{\alpha, \beta} \text{PS}(\delta_{\text{line}}(\alpha, \beta))$ à l'aide de leurs p-values respectives.
- **Comparaison des scores de prédiction :** Nous évaluons et comparons la précision des prédictions des modèles BADRDI, NP et PS-OP en utilisant les scores AUC-ROC pour la régression logistique et les scores C-index pour la régression de Cox, afin de déterminer quel modèle prédit le mieux les résultats de survie des patients.
- **Analyse du pouvoir discriminant :** Cette analyse examine la capacité de chaque modèle à différencier les résultats de survie des patients. Nous catégorisons les cohortes de patients en quartiles en fonction de $1 - \text{ADRDI}$, $\delta_{\text{line}}(\alpha = \frac{1}{2}, \beta = \frac{1}{2})$ et $\text{argmax}_{\alpha, \beta} \text{PS}(\delta_{\text{line}}(\alpha, \beta))$. En analysant les courbes de survie de Kaplan-Meier pour ces quartiles, nous évaluons l'efficacité des modèles à distinguer les différents résultats de survie, validée par des tests de log-rank pour la significativité statistique. La méthode est détaillée ci-dessous.

Notation - Code couleur des critères d'évaluation de l'analyse comparative

Dans la suite du chapitre, une couleur est attribuée à chaque critère d'évaluation de la performance des méthodes.

- : Significativité de la contribution de la variable explicative au modèle.
- : Comparaison des scores de prédiction.
- : Analyse du pouvoir discriminant.

La triple analyse comparative (BADRDI - NP - PS-OP) nécessite la réalisation complète des trois étapes de l'algorithme décrit dans la section 5.3.3.7. En revanche, le réalisation des deux premières étapes de l'algorithme suffisent pour la double analyse comparative (BADRDI - NP). Nous verrons dans les sections 5.7.3 et 5.8.2, que nous utilisons la double analyse comparative (BADRDI -NP) pour valider la méthode ProtoDrift sur deux hôpitaux.

5.4.1 Analyse du pouvoir discriminant

Le pouvoir discriminant d'une méthode se réfère à sa capacité à bien classer les différents résultats. Dans notre contexte, nous voulons comparer la capacité de BADRDI, NP et PS-OP à distinguer les différents résultats de survie des patients. Pour ce faire, nous divisons les cohortes de patients en quartiles en fonction de leurs valeurs de $1 - \text{ADRDI}$, $\delta_{\text{line}}(\alpha = \frac{1}{2}, \beta = \frac{1}{2})$ et $\text{argmax}_{\alpha, \beta} \text{PS}(\delta_{\text{line}}(\alpha, \beta))$ à la fin d'une ligne de traitement pour évaluer respectivement le pouvoir discriminant de BADRDI, NP et PS-OP. Pour rappel, l'ADRDI et la dissimilarité δ_{line} varient entre 0 et 1, mais une valeur élevée d'ADRDI signifie un suivi fidèle du traitement, tandis que c'est l'inverse pour δ_{line} . Pour aligner le sens des deux métriques, nous formons les quartiles avec le complément d'ADRDI (*i.e.*, $1 - \text{ADRDI}$) pour évaluer la méthode BADRDI. Nous considérons qu'il existe une différence significative de temps de survie entre deux quartiles si la p-value du test de log-rank est inférieure à 0,05.

5.5 Exploration de l'impact du temps par rapport à la dose sur la survie globale

Le résultat de l'algorithme de *grid search* nous fournit un score de prédiction pour chaque combinaison de paires (α, β) . Cela permet d'étudier la variabilité de l'association entre δ_{line} et la survie globale en fonction des valeurs de α et β . Par leurs définitions (5.16 et 5.17) :

- Une valeur élevée de α (proche de 1) signifie que $\omega_t > \omega_d$, c'est-à-dire que les écarts temporels par rapport au protocole prévu pour l'administration des médicaments anticancéreux ont plus d'impact sur la survie globale que les écarts de dosage.
- Une valeur élevée de β (proche de 1) signifie que $\omega_{inter} > \omega_{intra}$, c'est-à-dire que le retard de l'administration de l'ensemble du cycle de chimiothérapie a plus d'impact sur la survie globale que les écarts de calendrier et de dosage des administrations de médicaments anticancéreux au sein du cycle.

Pour une cohorte de patients, cette variabilité peut être étudiée en créant une carte de chaleur(*heatmap*) des valeurs de score de prédiction obtenues en fonction des valeurs de α et β .

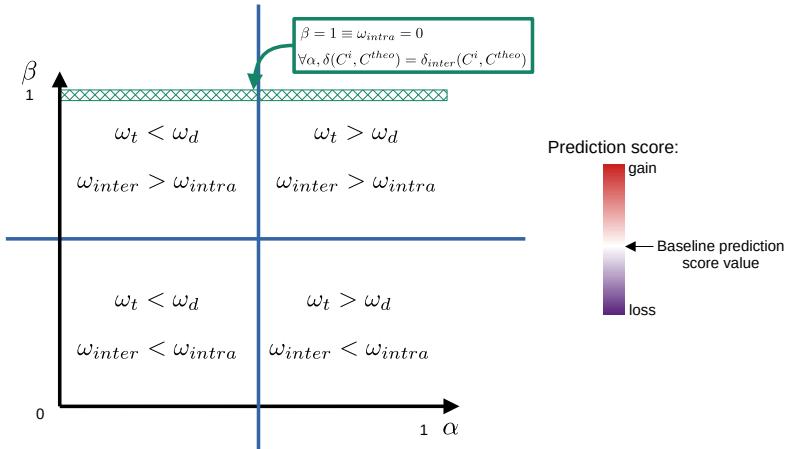


FIGURE 5.9 – Interprétation des profils de *heatmap*.

Comme illustré dans la figure 5.9, nous pouvons catégoriser les profils de *heatmap* en quatre zones :

- $\alpha \rightarrow 0$ et $\beta \rightarrow 0 \Leftrightarrow \omega_t < \omega_d$ et $\omega_{inter} < \omega_{intra}$: les écarts de dosage des administrations de médicaments anticancéreux ont plus d'impact sur la survie globale que leurs écarts de calendrier, que ce soit le calendrier de l'administration des médicaments au sein du cycle (pondéré par ω_t) ou la régularité de l'administration de l'ensemble du cycle (pondéré par ω_{inter}).
- $\alpha \rightarrow 1$ et $\beta \rightarrow 0 \Leftrightarrow \omega_t > \omega_d$ et $\omega_{inter} < \omega_{intra}$: les écarts de calendrier de l'administration des médicaments anticancéreux au sein du cycle ont plus d'impact sur

la survie globale que les écarts de dosage, et les écarts de calendrier et de dosage des administrations de médicaments anticancéreux (pondérés par ω_{intra}) ont plus d'impact sur la survie globale que la régularité de l'administration de l'ensemble du cycle (pondéré par ω_{inter}).

- $\alpha \rightarrow 0$ et $\beta \rightarrow 1 \Leftrightarrow \omega_t < \omega_d$ et $\omega_{\text{inter}} > \omega_{\text{intra}}$: les écarts de dosage des administrations de médicaments anticancéreux au sein du cycle ont plus d'impact sur la survie globale que les écarts de calendrier au sein du cycle, et la régularité de l'administration de l'ensemble du cycle a plus d'impact sur la survie globale que les écarts de calendrier et de dosage des administrations de médicaments au sein du cycle.
- $\alpha \rightarrow 1$ et $\beta \rightarrow 1 \Leftrightarrow \omega_t > \omega_d$ et $\omega_{\text{inter}} > \omega_{\text{intra}}$: les écarts de dosage des administrations de médicaments anticancéreux au sein du cycle ont plus d'impact sur la survie globale que les écarts de calendrier au sein du cycle, et la régularité de l'administration de l'ensemble du cycle a plus d'impact sur la survie globale que les écarts de calendrier et de dosage des administrations de médicaments au sein du cycle.

Propriété - ligne supérieure de la *heatmap*

En haut de la *heatmap*, lorsque $\beta = 1$, le score de prédiction devrait être constant (c'est-à-dire que la couleur devrait être uniforme).

Démonstration. En effet, pour toutes les valeurs de α , $\delta_{\text{line}}(\alpha, \beta = 1)$ est une constante. Comme les modèles ne varient qu'avec la variable explicative $\delta_{\text{line}}(\alpha, \beta)$, $\delta_{\text{line}}(\alpha, \beta = 1)$ étant une constante, les scores de prédiction réalisés avec cette dernière sont également constants.

$$\begin{aligned} \beta = 1 &\Leftrightarrow \frac{\omega_{\text{inter}}}{\omega_{\text{inter}} + \omega_{\text{intra}}} = 1 \\ &\Leftrightarrow \omega_{\text{intra}} = 0 \\ &\Leftrightarrow \delta_{\text{cycle}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}}) = \delta_{\text{inter}} \\ &\Leftrightarrow \delta_{\text{line}}(\mathcal{L}, \mathcal{C}^{\text{theo}}) = \frac{1}{l} \sum_{i=1}^l \delta_{\text{inter}}(\mathcal{C}^i, \mathcal{C}^{\text{theo}}) \\ &\Leftrightarrow \forall \alpha, \delta_{\text{line}}(\alpha, \beta = 1) \text{ est constant} \end{aligned}$$

□

5.6 Bilan sur les méthodes proposées et leurs évaluations

À ce stade, nous avons détaillé deux méthodes autour de ProtoDrift. Ces méthodes permettent de :

1. Calculer les dissimilarités de la méthode ProtoDrift (*cf. section 5.2*). La projet gitlab ProtoDrift (accessible sur <https://gitlab.inria.fr/arogier/protodrift>)

propose une implémentation pour obtenir une table de dissimilarités à partir d'une table ou d'un graphe **ChemoOnto**. Un *pipeline* ChemoOnto-ProtoDrift présent dans le projet **ChemoOnto** (accessible sur <https://gitlab.inria.fr/arogier/ChemoOntoTox>), permet de tester le code sur les données de cinq faux patients ;

2. Associer les dissimilarités ProtoDrift ligne-protocole δ_{line} aux survies des patients, et optimiser les poids de ProtoDrift sur les scores de prédiction de la survie globale. (*cf.* section 5.3)

Nous avons ensuite détaillé deux méthodes qui évaluent et analysent les méthodes développées autour de ProtoDrift. Ces méthodes permettent de :

- a) Comparer les prédictions de survies globales des différentes méthodes pour mesurer l'adhésion aux chimiothérapies (*cf.* section 5.4) :
 - La méthode standard utilisant la dose-intensité relative de tous les anticancéreux sur une ligne de traitement (ADRD), qui nous sert de baseline. On l'a nommée la méthode "BADRD".
 - La méthode utilisant un ProtoDrift naïf, sans a priori sur les valeurs des poids associés aux différentes dissimilarités de ProtoDrift. On l'a nommée la méthode "NP".
 - La méthode utilisant un ProtoDrift avec des poids associés aux dissimilarité optimisé pour mieux prédire la survie globale. On l'a nommée la méthode "PS-OP".
- b) Explorer l'impact relatif des poids associés aux dissimilarités sur la survie globale (*cf.* section 5.5)

Le projet gitlab **ProtoDrift-Surv** (accessible sur <https://gitlab.inria.fr/arogier/protodrift-surv>) permet l'application de la méthode 2, la méthode d'évaluation a) et d'analyse b). Cependant ces implémentations nécessitent des données cliniques personnelles sur les patients et ne sont pas testables sur un jeu de données de petite taille. Nous ne fournissons pas de jeu de données test dans ce projet.

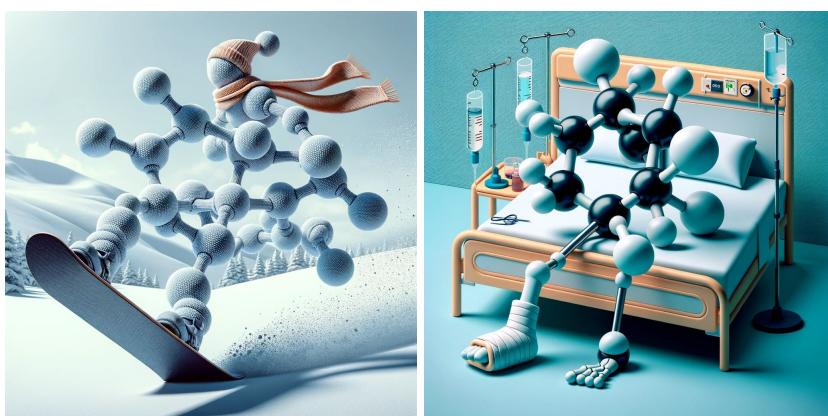


FIGURE 5.10 – Rares images d'un ProtoDrift. Gauche : avatar de **ProtoDrift**. Comme nous le verrons, en général, plus la dérive est forte, plus les chances de survie sont faibles. À droite, avatar de **ProtoDrift-Surv**, ce dernier paie les conséquences d'un dérapage incontrôlé.

Dans la suite du chapitre, nous fournissons les résultats de l'application de ces méthodes sur les données de deux hôpitaux, l'HEGP et le CHU de Bordeaux.

5.7 Application de ProtoDrift sur les données l'HEGP et du CHU de Bordeaux

5.7.1 Sources de données et conception de l'étude

Nous avons collecté de manière indépendante les schémas thérapeutique théoriques et leurs suivis réels par les patients dans deux hôpitaux français, l'Hôpital Européen Georges Pompidou (HEGP) et le Centre Hospitalier Universitaire de Bordeaux (CHU de Bordeaux), du 1er juillet 2003 au 15 décembre 2021 (environ 18,5 ans) (*cf. sections 1.4.5 et 1.4.6*). Nous nous sommes appuyés sur ChemoOnto pour reconstituer le parcours de traitement de chimiothérapie à partir des enregistrements de médicaments anticancéreux et des schémas thérapeutiques théoriques (*cf. chapitre 4*) [170, 105]. Les données sur la survie des patients, pendant et après leur séjour, sont incluses dans les deux hôpitaux et utilisées comme résultat dans cette étude (*cf. section 1.4.8*) [41].

Les groupes de patients ont été définis par localisation de la tumeur, mais seuls les groupes de plus de 400 patients ont été pris en compte. Pour chaque localisation, deux cohortes sont créées, pour la première et la deuxième ligne de traitement suivie. Une description des cohortes est présentée dans les tables 5.12 et 5.13. Bien que les valeurs moyennes de l'ADRDI et de la δ_{line} de NP soient similaires dans de nombreuses cohortes, leurs distributions distinctes suggèrent que chaque métrique capture différents aspects de l'impact du traitement.

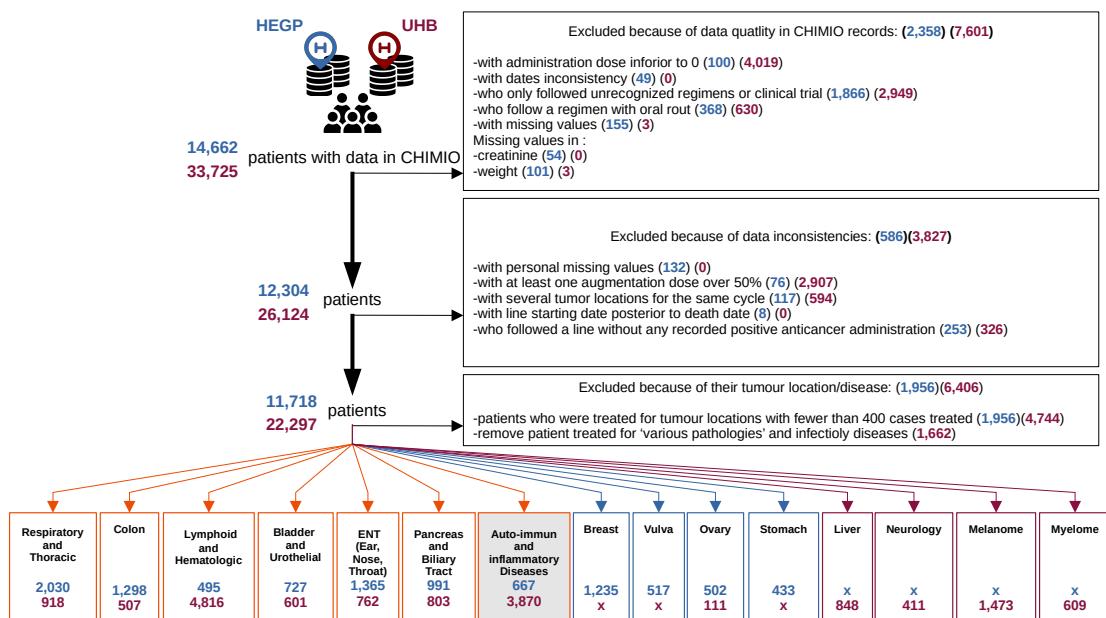


FIGURE 5.11 – Conception de l'étude. La couleur bleue est associée à l'HEGP et la couleur bordeaux au CHU de Bordeaux. Les cohortes spécifiques à l'HEGP sont donc encadrées en bleu, celles spécifiques au CHU de Bordeaux sont encadrées en bordeaux, et les cohortes communes aux deux hôpitaux sont encadrées en orange.

Nous avons exclu 2 944 patients à l'HEGP et 11 428 patients au CHU de Bordeaux en raison de la qualité ou de l'incohérence des données. Les détails sur la sélection des patients sont fournis dans la figure 5.11 et le paragraphe suivant. Onze groupes de maladies comprenant plus de 400 patients ont été conservés à l'HEGP et au CHU de Bordeaux, incluant respectivement 9 762 et 15 891 patients. Ainsi, l'étude est composée de 22 cohortes de l'HEGP et de 22 cohortes du CHU de Bordeaux, décrites dans les tableaux 5.12 et 5.13. Notons que les groupes de maladies ne sont pas exclusifs, et que chaque patient d'un groupe de maladies a suivi une première ligne de traitement, mais pas nécessairement une deuxième ligne. Ainsi, chaque groupe de maladies a la même taille que sa cohorte de première ligne correspondante, et à une taille supérieure à celle de sa cohorte de deuxième ligne. Parmi les 15 groupes de maladies, sept sont communs à l'HEGP et au CHU de Bordeaux. Le groupe des maladies auto-immunes et inflammatoires est particulier car il fait référence à un groupe de patients qui ne sont pas atteints d'un cancer mais de maladies auto-immunes et inflammatoires (par exemple, polyarthrite rhumatoïde, maladie de Wegener et lupus). Ces patients suivent un traitement cyclique qui a une organisation similaire aux traitements de chimiothérapies, avec notamment des administration d'anti-corps monoclonaux (cf. 1.1.3). Leurs données de suivie sont donc saisies dans le logiciel Chimio®. (cf. . 1.4.6)

Notation - Code couleur de la provenance des jeux de données

Dans la suite du chapitre, une couleur est attribuée aux résultats obtenus avec chaque hôpital.

- : Données provenant de l'HEGP
- : Données provenant du CHU de Bordeaux

5.7.1.1 Détails sur les processus d'exclusion

Parmi les patients disponibles dans le logiciel Chimio®, nous avons exclu ceux pour lesquels il est impossible de calculer ProtoDrift en raison de valeurs manquantes. Nous avons également exclu les patients pour lesquels il y a une incohérence de date dans les données, comme une administration datée en dehors des dates de cycle rapportées, des administrations enregistrées après la date de décès et les patients sans aucune administration de médicament anti-cancéreux. Il existe un nombre de protocoles administrés pour lesquels nous ne parvenons pas à trouver le protocole théorique correspondant. Nous avons exclu les patients qui ont suivi uniquement des protocoles non reconnus. Ensuite, nous avons exclu les patients avec un sexe ou une date de naissance manquant, car ces données sont nécessaires pour effectuer notre analyse de survie. Nous avons également exclu les patients qui ont suivi des protocoles avec capécitabine ou amifostine, car les doses entre le protocole théorique et l'administration étaient systématiquement trop différentes. Nous supposons qu'il doit y avoir une erreur dans la description des protocoles théoriques associés. Une représentation schématique de différentes étapes d'exclusion des patients est présentée figure 5.11.

	Demographic characteristics				Overall Survival				Metric			
	Sex Count (%)		Age at line start		Mortality count (%)		Mean		SD		Distribution	
	M	F	Mean	SD	3 years	5 years	δ_{line}	ADRDI	δ_{line}	ADRDI	δ_{line}	ADRDI
Respiratory and Thoracic												
First line (2030)	1339 (66%)	691 (34%)	64.62	11.61	1442 (71%)	1550 (76.4%)	0.14	0.82	0.11	0.18		
First line (918)	587 (63.9%)	331 (36.1%)	56.74	20.26	516 (56.2%)	548 (59.7%)	0.17	0.73	0.12	0.27		
Second line (850)	541 (63.6%)	309 (36.4%)	63.89	11.05	645 (75.9%)	689 (81.1%)	0.14	0.79	0.13	0.21		
Second line (346)	215 (62.1%)	131 (37.9%)	59.78	13.22	213 (61.6%)	229 (66.2%)	0.2	0.67	0.14	0.27		
Colon												
First line (1298)	730 (56.2%)	568 (43.8%)	65.37	13.14	720 (55.5%)	836 (64.4%)	0.13	0.78	0.1	0.2		
First line (507)	304 (60%)	203 (40%)	65.33	12.45	295 (58.2%)	316 (62.3%)	0.15	0.61	0.13	0.24		
Second line (752)	407 (54.1%)	345 (45.9%)	65.82	13.14	487 (64.8%)	558 (74.2%)	0.13	0.75	0.1	0.2		
Second line (216)	132 (61.1%)	84 (38.9%)	64.9	13.07	143 (66.2%)	150 (69.4%)	0.21	0.54	0.18	0.33		
Lymphoid and Hematologic												
First line (495)	286 (57.8%)	209 (42.2%)	64.77	21.11	219 (44.2%)	245 (49.5%)	0.16	0.77	0.12	0.23		
First line (4816)	2767 (57.5%)	2049 (42.5%)	61.5	17.09	1455 (30.2%)	1622 (33.7%)	0.25	0.54	0.14	0.3		
Second line (232)	140 (60.3%)	92 (39.7%)	63.28	19.88	101 (43.5%)	118 (50.9%)	0.19	0.69	0.15	0.29		
Second line (2741)	1575 (57.5%)	1166 (42.5%)	61.42	16.47	724 (26.4%)	812 (29.6%)	0.25	0.52	0.13	0.28		
Bladder and Urothelial												
First line (727)	563 (77.4%)	164 (22.6%)	69.14	11.89	419 (57.6%)	471 (64.8%)	0.13	0.8	0.11	0.23		
First line (601)	485 (80.7%)	116 (19.3%)	66.9	11.34	144 (24%)	160 (26.6%)	0.15	0.83	0.12	0.27		
Second line (222)	174 (78.4%)	48 (21.6%)	67.27	13.07	163 (73.4%)	171 (77%)	0.13	0.79	0.12	0.23		
Second line (131)	109 (83.2%)	22 (16.8%)	64.6	12.95	44 (33.6%)	46 (35.1%)	0.14	0.78	0.15	0.29		
ENT (Ear, Nose, Throat)												
First line (1365)	1096 (80.3%)	269 (19.7%)	62.67	11.42	736 (53.9%)	836 (61.2%)	0.13	0.84	0.13	0.19		
First line (762)	576 (75.6%)	186 (24.4%)	59.11	12.96	305 (40%)	336 (44.1%)	0.09	0.92	0.08	0.18		
Second line (750)	610 (81.3%)	140 (18.7%)	63.08	11.13	492 (65.6%)	532 (70.9%)	0.13	0.82	0.14	0.22		
Second line (294)	229 (77.9%)	65 (22.1%)	59.44	13.06	163 (55.4%)	173 (58.8%)	0.09	0.86	0.1	0.22		
Pancreas and Biliary Tract												
First line (991)	540 (54.5%)	451 (45.5%)	68.34	11.52	741 (74.8%)	786 (79.3%)	0.19	0.76	0.12	0.19		
First line (803)	461 (57.4%)	342 (42.6%)	67.91	11	622 (77.5%)	640 (79.7%)	0.15	0.74	0.09	0.2		
Second line (510)	268 (52.5%)	242 (47.5%)	67.68	11.4	408 (80%)	426 (83.5%)	0.18	0.69	0.12	0.21		
Second line (271)	152 (56.1%)	119 (43.9%)	66.61	11.7	209 (77.1%)	212 (78.2%)	0.19	0.64	0.12	0.24		
Autoimmune and Inflammatory												
First line (667)	342 (51.3%)	325 (48.7%)	48.52	20.62	108 (16.2%)	127 (19%)	0.12	0.9	0.11	0.19		
First line (3870)	1738 (44.9%)	2132 (55.1%)	50.76	19.46	304 (7.9%)	382 (9.9%)	0.23	0.64	0.14	0.33		
Second line (377)	207 (54.9%)	170 (45.1%)	47.18	20.26	44 (11.7%)	51 (13.5%)	0.11	0.92	0.11	0.17		
Second line (1612)	683 (42.4%)	929 (57.6%)	50.18	18.51	94 (5.8%)	125 (7.8%)	0.25	0.57	0.15	0.36		
Breast												
First line (1235)	51 (4.1%)	1184 (95.9%)	57.27	13.71	398 (32.2%)	486 (39.4%)	0.09	0.89	0.09	0.17		
Second line (570)	23 (4%)	547 (96%)	57.36	13.11	263 (46.1%)	302 (53%)	0.1	0.84	0.11	0.19		

FIGURE 5.12 – Caractéristiques démographiques, de survie globale et des métriques statistiques qualitatives des cohortes incluses dans l'analyse (1/2)

	Demographic characteristics				Overall Survival				Metric			
	Sex Count (%)		Age at line start		Mortality count (%)		Mean		SD		Distribution	
	M	F	Mean	SD	3 years	5 years	δ_{line}	ADRDI	δ_{line}	ADRDI	δ_{line}	ADRDI
Vulva												
First line (517)	16 (3.1%)	501 (96.9%)	63.57	13.15	295 (57.1%)	330 (63.8%)	0.15	0.82	0.14	0.19		
Second line (242)	4 (1.7%)	238 (98.3%)	64.39	12.13	164 (67.8%)	187 (77.3%)	0.17	0.81	0.15	0.2		
Ovary												
First line (502)	3 (0.6%)	499 (99.4%)	65.34	13.13	251 (50%)	291 (58%)	0.17	0.84	0.14	0.15		
Second line (321)	1 (0.3%)	320 (99.7%)	66.27	11.64	173 (53.9%)	205 (63.9%)	0.12	0.85	0.11	0.17		
Stomach												
First line (433)	279 (64.4%)	154 (35.6%)	62.73	13.5	288 (66.5%)	302 (69.7%)	0.13	0.77	0.09	0.2		
Second line (207)	135 (65.2%)	72 (34.8%)	61.75	13.1	150 (72.5%)	156 (75.4%)	0.13	0.76	0.1	0.21		
Liver												
First line (848)	723 (85.3%)	125 (14.7%)	68.3	10.1	391 (46.1%)	451 (53.2%)	0.22	0.84	0.11	0.28		
Second line (152)	131 (86.2%)	21 (13.8%)	70.47	9.68	74 (48.7%)	80 (52.6%)	0.21	0.64	0.17	0.36		
Neurology												
Second line (411)	198 (48.2%)	213 (51.8%)	54.8	16.01	27 (6.6%)	28 (6.8%)	0.27	0.46	0.17	0.41		
First line (685)	317 (46.3%)	368 (53.7%)	54.65	17.23	52 (7.6%)	66 (9.6%)	0.27	0.56	0.13	0.41		
Melanoma												
First line (1473)	882 (59.9%)	591 (40.1%)	67.04	14.57	689 (46.8%)	731 (49.6%)	0.13	0.95	0.14	0.11		
Second line (628)	371 (59.1%)	257 (40.9%)	65.84	14.07	274 (43.6%)	295 (47%)	0.1	0.93	0.14	0.14		
Myeloma												
First line (609)	328 (53.9%)	281 (46.1%)	67.22	10.84	162 (26.6%)	189 (31%)	0.18	0.65	0.13	0.27		
Second line (389)	217 (55.8%)	172 (44.2%)	68.43	10.77	105 (27%)	113 (29%)	0.21	0.54	0.13	0.26		
All locations/diseases												
First line (9762) First line (15891)	5046 (51.7%) 9167 (57.7%)	4716 (48.3%) 6724 (42.3%)	63.13 59.94	14.49 17.61	5351 (54.8%) 4934 (31%)	5944 (60.9%) 5440 (34.2%)	0.14 0.21	0.82 0.68	0.12 0.14	0.19 0.32		
Second line (4910) Second line (7189)	2464 (50.2%) 4011 (55.8%)	2446 (49.8%) 3178 (44.2%)	62.71 59.68	14.26 16.87	3019 (61.5%) 2069 (28.8%)	3311 (67.4%) 2262 (31.5%)	0.13 0.22	0.79 0.6	0.12 0.15	0.22 0.33		

FIGURE 5.13 – Caractéristiques démographiques, de survie globale et des métriques statistiques qualitatives des cohortes incluses dans l'analyse (2/2)

5.7.2 Correspondance entre les noms de localisation de cancer à l'HEGP et au CHU de Bordeaux

Sans surprise, les noms donnés aux localisations/maladies ne sont pas identiques entre l'HEGP et le CHU de Bordeaux. Le CHU de Bordeaux est plus précis, et souvent un nom de localisation de tumeur à l'HEGP regroupe plusieurs noms au CHU de Bordeaux. La Docteure Mathilde Pezot, interne en rhumatologie, nous a aidé à aligner les noms de localisations/maladies. Le tableau 5.3 détaille la correspondance réalisée pour les sept localisations communes à l'HEGP et au CHU de Bordeaux sélectionnées pour l'étude (*cf. figure 5.11*). Les appellation conservées pour notre étude sont celles de l'HEGP.

HEGP	CHU de Bordeaux
RESPIRATOIRE et THORACIQUE	PNEUMO - POUMON NON A PETITES CELLULE, PNEUMO - POUMON PETITES CELLULES, CARDIOLOGIE, PNEUMO
ORL	OTO-RHINO-LARYNGO - PHARYNX, OTO-RHINO-LARYNGO - LARYNX, OTO-RHINO-LARYNGO - BOUCHE, OTO-RHINO-LARYNGO, OTO-RHINO-LARYNGO - GLANDES SALIVAIRES
COLON	GASTRO-ENTERO - COLON
PANCRÉAS et VOIES BILIAIRES	GASTRO-ENTERO - PANCRÉAS, GASTRO-ENTERO - VOIES BILIAIRES
VESSIE et UROTHÉLIAL	NÉPHRO-URO - VESSIE, NÉPHRO-URO - VOIES URINAIRES
AUTO-IMMUN et INFLAMMATOIRE	MALADIE DE CROHN, maladies AUTO-IMMUNES, POLYARTHRITE RHUMATOÏDE, RECTOCOLITE HÉMORRAGIQUE, VASCULARITE
LYMPHOÏDE et HÉMATOLOGIQUE	HÉMATO ADULTE - LAM, HÉMATOLOGIE, HÉMATO ADULTE - LYMPHOMES NON HODGKINIENS, HÉMATO ADULTE - LYMPHOMES HODGKINIENS, HÉMATO ADULTE - LLC, DERMATO - LYMPHOME, HÉMATO ADULTE - LAL, HÉMATO ADULTE ALLOGREFFE, NEUTROPÉNIES CYTOTOXIQUES

TABLE 5.3 – Correspondance des noms de localisations de cancers/maladies entre HEGP et CHU de Bordeaux. Les appellation conservées pour notre étude sont celles de l'HEGP.

5.7.3 Évaluation de ProtoDrift sur deux jeux de données indépendants

Les objectifs de l'évaluation sur deux jeux de données indépendants visent à démontrer :

- la faisabilité de la méthode
- la validation de la métrique ProtoDrift

Nous nous attendons à ce que ProtoDrift prédisse globalement mieux la survie globale que la dose-intensité relative dans les deux hôpitaux. Pour tester notre hypothèse, nous effectuons la double analyse comparative décrite précédemment dans la section 5.4 et comparons les gains/pertes obtenus dans les deux hôpitaux. Pour des raisons de temps, l'optimisation des poids de ProtoDrift ne peut pas être réalisée pour toutes les cohortes du CHU de Bordeaux. Pour valider ProtoDrift, nous comparons les gains obtenus dans les deux hôpitaux entre les méthodes NP (Naive ProtoDrift) et BADRDI (Baseline All Drugs Relative Dose-Intensity). Pour valider notre hypothèse, nous devons obtenir une majorité de gains dans les cohortes testées dans les deux hôpitaux.

5.7.4 Configuration des paramètres d'application

Dans cette section, nous détaillons les paramètres d'application décrits dans la section 5.3.3.5 pour les analyses réalisées à l'HEGP et au CHU de Bordeaux.

À l'HEGP, nous avons réalisé les trois étapes de l'algorithme d'optimisation décrites dans la section 5.3.3.7 sur les 22 cohortes (les 11 localisations de cancer/maladies de la figure 5.11 et deux premières lignes de traitement) en prédisant la survie à 3 et à 5 ans avec les deux modèles de régression.

Ainsi, 88 triple analyses comparatives BADRDI-NP-OP (*cf. section 5.4*) et explorations de l'impact des poids relatifs sur la survie (*cf. section 5.5*) ont été réalisées (88 = 22 cohortes * 2 temps de survie * 2 modèles de régression).

L'intégralité des résultats peut être visualisée en naviguant sur le site <https://files.inria.fr/protodrift-surv/>.

5.7.4.1 Paramètres d'application de l'algorithme d'optimisation des poids de ProtoDrift à l'HEGP

Cette section liste les valeurs des paramètres personnalisables de l'algorithme d'optimisation détaillés dans la section 5.3.3.7.

- Les paramètres d'adaptation
 - **Localisation de la tumeur/maladies** : respiratoire et thoracique, côlon, estomac, lymphoïde et hématologique, vessie et urothélial, ovaire, sein, vulve, ORL, pancréas et voies biliaires, auto-immun et inflammatoire (pas une localisation de cancer), toutes localisations confondues
 - **Numéro de ligne** : 1 et 2
- Paramètre de résultat de l'optimisation
 - **Temps de survie** : survie à 3 et 5 ans
- Paramètres de l'algorithme

- **Pas du grid search** : 0.1 (121 combinaisons $\alpha, \beta \in [0 : 1]$)
- **Nombre d'échantillons bootstrap** : 500
- **Pourcentage de division** : 0.7
- Paramètre de stratification
 - **Tranches d'âge** : 5 ans

Pour accélérer le processus, nous avons parallélisé sur 20 processeurs les étapes 1 et 2 de l'algorithme (*cf.* figure 5.7) sur les échantillons de *bootstrap* et l'étape 3 (*cf.* figure 5.8) sur les combinaisons α, β .

Au CHU de Bordeaux, seules les deux premières étapes de l'algorithme ont été réalisées sur les 22 cohortes en prédisant la survie à 5 ans avec les deux modèles de régression. La dernière étape, la recherche par grille qui permet l'optimisation, a été réalisée pour la cohorte de patients atteints de cancer respiratoire et thoracique suivie en première ligne de traitement. Ainsi, nous avons réalisé 44 doubles analyses comparatives (BADRDI-NP) et une triple analyse comparative et explorations de l'impact des poids relatifs sur la survie.

5.7.4.2 Paramètres d'application de l'algorithme d'optimisation des poids de ProtoDrift au CHU de Bordeaux

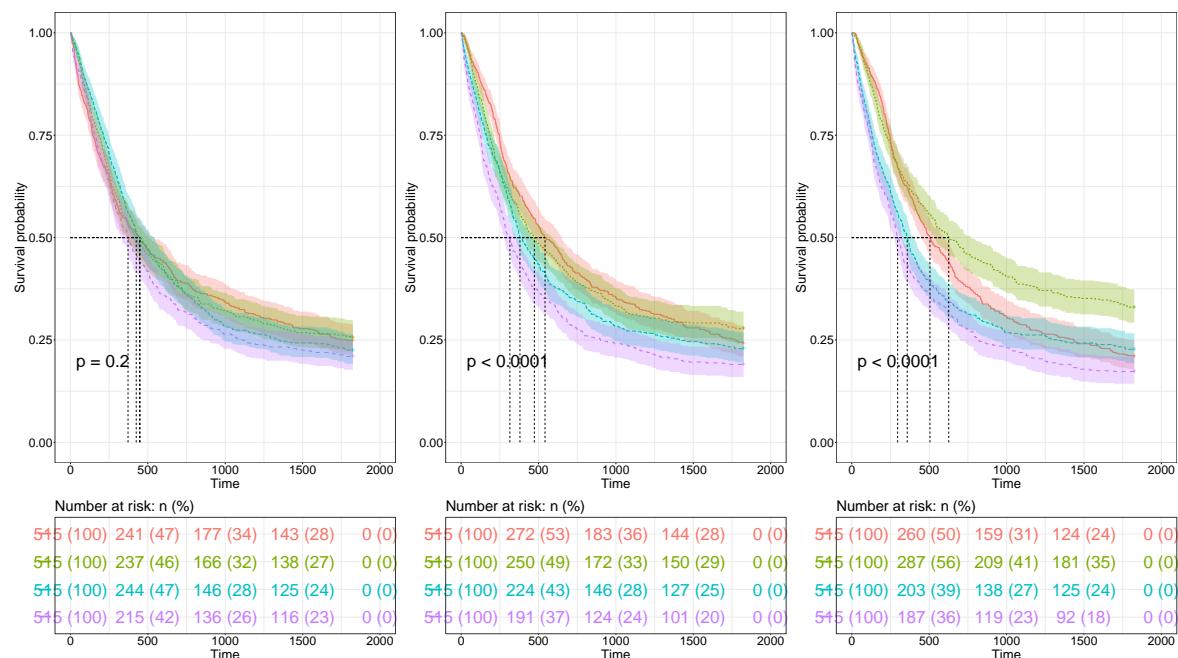
- Les paramètres d'adaptation
 - **Localisation de la tumeur/maladies** : respiratoire et thoracique, côlon, lymphoïde et hématologique, vessie et urothelial, ORL, pancréas et voies biliaires, auto-immun et inflammatoire (pas une localisation de cancer), foie, neurologie, mélanome, myélome, toutes localisations confondues
 - **Numéro de ligne** : 1 et 2
- Paramètre de résultat de l'optimisation
 - **Temps de survie** : survie à 5 ans
- Paramètres de l'algorithme
 - **Pas du grid search seulement pour les cancers respiratoires et thoraciques, en première ligne** : 0.1 (121 combinaisons $\alpha, \beta \in [0 : 1]$)
 - **Nombre d'échantillons bootstrap** : 500
 - **Pourcentage de division** : 0.7
- Paramètre de stratification
 - **Tranches d'âge** : 5 ans

5.8 Résultats d'application

5.8.1 Performances prédictives de la survie globale à 5 ans de la cohorte respiratoire et thoracique en première ligne

À l'HEGP, les poids de ProtoDrift ont été optimisés pour toutes les cohortes, et les résultats complets sont disponibles sur <https://files.inria.fr/protodrift-surv/>.

Ici, nous nous concentrons sur la triple analyse comparative BADRDI-NP-C-index-OP obtenue pour les cohortes respiratoire et thoracique en première ligne de traitement (R/T) de l'HEGP et du CHU de Bordeaux.



Quartiles:
 + first
 + second
 + third
 + fourth

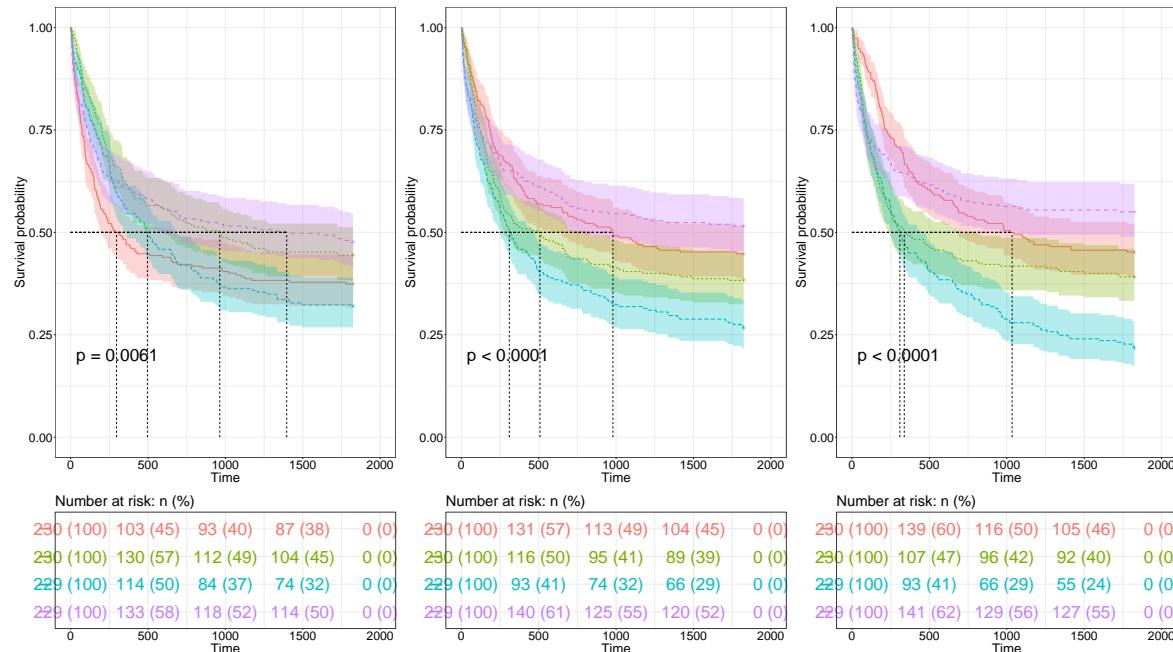
FIGURE 5.14 – Courbes de Kaplan-Meier construites sur les quartiles de 1 – ADRDI (gauche), δ_{line} de NP (milieu) et C-index- δ_{line} (droite) à la fin de la première ligne de la cohorte respiratoire et thoracique (R/T) de l'HEGP.

Cox regression models comparison													
Model fit				Model prediction with C-index				Pairwise quartile LogRank p-value					
	θ_1	CI	sd	p-value	Value	CI	sd	1 st vs 2 nd	1 st vs 3 rd	2 nd vs 3 rd	1 st vs 4 th	2 vs 4	3 rd vs 4 th
OP $\alpha=1 \beta=0$	1.101	(0.72 - 1.49)	0.225	0	0.551	(0.52 - 0.58)	0.019	0.001	0.042	0	0	0	0.034
NP $\alpha=0.5 \beta=0.5$	1.442	(0.76 - 2.10)	0.398	0.006	0.538	(0.51 - 0.57)	0.019	0.893	0.049	0.06	0	0	0.033
BADRDI	-0.283	(-0.62 - 0.07)	0.212	0.234	0.519	(0.49 - 0.55)	0.019	0.888	0.888	0.209	0.209	0.209	0.209

FIGURE 5.15 – Tableau de résultats de l'analyse comparative des modèles de régression de Cox pour BADRDI, NP et C-index-OP sur les performances prédictives de la survie globale à 5 ans pour la cohorte de première ligne R/T de l'HEGP

Dans l'analyse comparative de la cohorte respiratoire et thoracique en première ligne de l'HEGP, les modèles de Cox NP et C-index-OP montrent des performances supérieures selon nos trois critères clés par rapport au modèle BADRDI. Tout d'abord, NP et C-index-OP montrent des contributions significatives dans les modèles de régression de Cox, indiquant une prédiction fiable du risque de mortalité, ce qui n'est pas observé avec BADRDI (cf.

partie orange de la table 5.15). Ensuite, les deux modèles surpassent BADRDI en précision prédictive, avec des scores de C-index plus élevés (*cf.* partie bleue de la table 5.15). Enfin, l'analyse du pouvoir discriminant, utilisant les courbes de survie de Kaplan-Meier, confirme que NP et C-index-OP différencient efficacement les résultats de survie des patients entre les quartiles (*cf.* figure 5.14), ce qui est validé par des résultats de tests log-rank significatifs (*cf.* partie rose de la table 5.15).



Quartiles:
+ first
+ second
+ third
+ fourth

FIGURE 5.16 – Courbes de Kaplan-Meier construites sur les quartiles de 1 – ADRDI (gauche), δ_{line} de NP (milieu) et C-index- δ_{line} (droite) à la fin de la première ligne de la cohorte respiratoire et thoracique (R/T) du CHU de Bordeaux.

Cox regression models comparison													
Model fit				Model prediction with C-index				Pairwise quartile LogRank p-value					
	θ_1	CI	sd	p-value	Value	CI	sd	1 st vs 2 nd	1 st vs 3 rd	2 nd vs 3 rd	1 st vs 4 th	2 vs 4	3 rd vs 4 th
OP $\alpha=1 \beta=0$	0.789	(0.22 - 1.34)	0.347	0.072	0.542	(0.49 - 0.60)	0.032	0.014	0	0.006	0.337	0.006	0
NP $\alpha=0.5 \beta=0.5$	1.152	(0.10 - 2.22)	0.674	0.155	0.534	(0.48 - 0.59)	0.031	0.085	0	0.024	0.401	0.024	0
BADRDI	-0.191	(-0.68 - 0.34)	0.316	0.375	0.527	(0.48 - 0.58)	0.031	0.022	0.982	0.022	0.032	0.982	0.022

FIGURE 5.17 – Tableau de résultats de l'analyse comparative des modèles de régression de Cox pour BADRDI, NP et C-index-OP sur les performances prédictives de la survie globale à 5 ans pour la cohorte de première ligne R/T du CHU de Bordeaux.

Dans l'analyse comparative du CHU de Bordeaux, bien qu'aucune des métriques étu-

diées ne montre de contribution significative dans les modèles de régression de Cox, les modèles NP et C-index-OP continuent de surpasser le modèle BADRDI en termes de prédictions globales avec le C-index (*cf.* partie bleue de la table 5.17). Nous pouvons observer qualitativement que les courbes de Kaplan-Meier sont plus espacées avec les modèles NP et C-index-OP qu'avec le modèle BADRDI, soulignant un meilleur pouvoir discriminant (*cf.* figure 5.16). Ces résultats valident la pertinence de l'utilisation de ProtoDrift avec la cohorte respiratoire et thoracique en première ligne de traitement.

Dans la suite, nous vérifions si ce résultat est généralisable à toutes les cohortes étudiées.

5.8.2 Évaluation de ProtoDrift

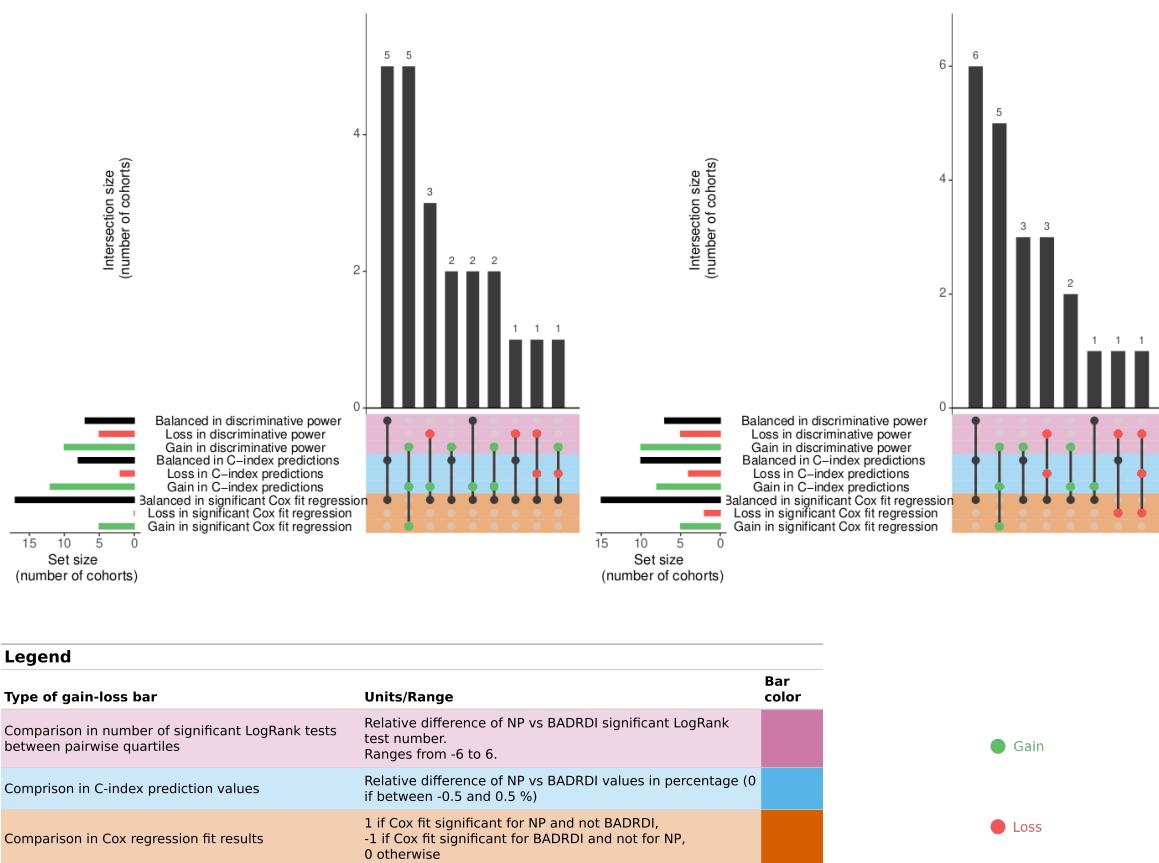


FIGURE 5.18 – Distribution des résultats de performances des 22 cohortes de l'HEGP (à gauche) et des 22 cohortes du CHU de Bordeaux (à droite). Les barres verticales quantifient le nombre de cohortes alignées sur des combinaisons spécifiques de résultats de performance. Ces combinaisons de performances sont illustrées dans la matrice codée par couleur en-dessous. La couleur des points indique la performance relative : vert pour les gains, rouge pour les pertes et noir pour les résultats équilibrés. Les barres horizontales montrent les nombres agrégés de cohortes connaissant des gains, des pertes ou des résultats équilibrés.

Les figures 5.18 et 5.19 confrontent la synthèse des résultats des doubles analyses comparatives NP-BADRDI des cohortes de l'HEGP et du CHU de Bordeaux. Elles révèlent que la méthode NP surpassé généralement la méthode BADRDI.

À l'HEGP, d'après la figure 5.18, 15 cohortes sur 22 (68% des cohortes) montrent au moins un type de gain de performance avec la méthode NP.

La figure 5.19 montre qu'il s'agit des cohortes Respiratoire et Thoracique (R/T), Colon (C), Lymphoïde et Hématologique (L/H), Vessie et Urothélial (V/U), Pancréas et Voies Biliaires (P/B), Maladies Auto-immunes et Inflammatoires (A/I), Sein (S) et Ovaies en première ligne, et Respiratoire et Thoracique (R/T), Colon (C), Vessie et Urothélial (V/U), ORL, Sein (S), Ovaies (O) et Estomac (E) en deuxième ligne.

Notamment, cinq cohortes montrent des gains dans les trois dimensions. Il s'agit des cohortes Respiratoire et Thoracique (R/T), Auto-immunes et Inflammatoires (A/I) et Ovaies (O) en première ligne et ORL en deuxième ligne.

Ces résultats suggèrent que ProtoDrift, même sans optimisation, offre une compréhension plus nuancée de l'adhésion au traitement par rapport à BADRDI.



FIGURE 5.19 – Trois évaluations des performances relatives de NP comparées à BADRDI sur les 22 cohortes de l'HEGP (à gauche) et les 22 cohortes du CHU de Bordeaux (à droite). Pour chaque cohorte, les barres à droite indiquent un gain, les barres à gauche indiquent une perte, et l'absence de barre indique un résultat équilibré dans la métrique d'évaluation correspondante.

La double analyse comparative à travers les cohortes du CHU de Bordeaux confirme l'efficacité de la méthode NP, avec une majorité de cohortes montrant des tendances d'amélioration similaires.

En effet, d'après la figure 5.18 11 cohortes sur 22 (50%) présentent au moins un gain de performance, et surpassent significativement la méthode BADRDI.

La figure 5.19 montre qu'il s'agit des cohortes Respiratoire et Thoracique (R/T), Colon

(C), Lymphoïde et Hématologique (L/H), ORL et Mélanome (M) en première ligne, et Colon (C), Lymphoïde et Hématologique (L/H), Vessie et Urothélial (V/U), ORL, Neurologie (N) et Mélanome (M) en deuxième ligne.

La méthode BADRDI surpassé NP dans cinq cohortes (Vessie et Urothélial (V/U), Foie (F), Neurologie (N), Myélome (My) en première ligne, Myélome (My) en deuxième ligne.

Cinq cohortes montrent également des gains dans les trois dimensions évaluées (Colon (C), ORL, Mélanome (M) en première ligne, Colon (C) et Mélanome (M) en deuxième ligne).

De façon générale, les résultats observés sur les deux jeux de données réels et indépendants sont cohérents. Cela confirme que ProtoDrift est globalement plus performant que l'approche standard RDI, même dans sa forme naïve. Ces résultats prometteurs suggèrent que l'optimisation de telles métriques pourrait améliorer davantage notre compréhension de l'adhésion au protocole et de ses implications.

5.8.3 Optimisation des poids de ProtoDrift

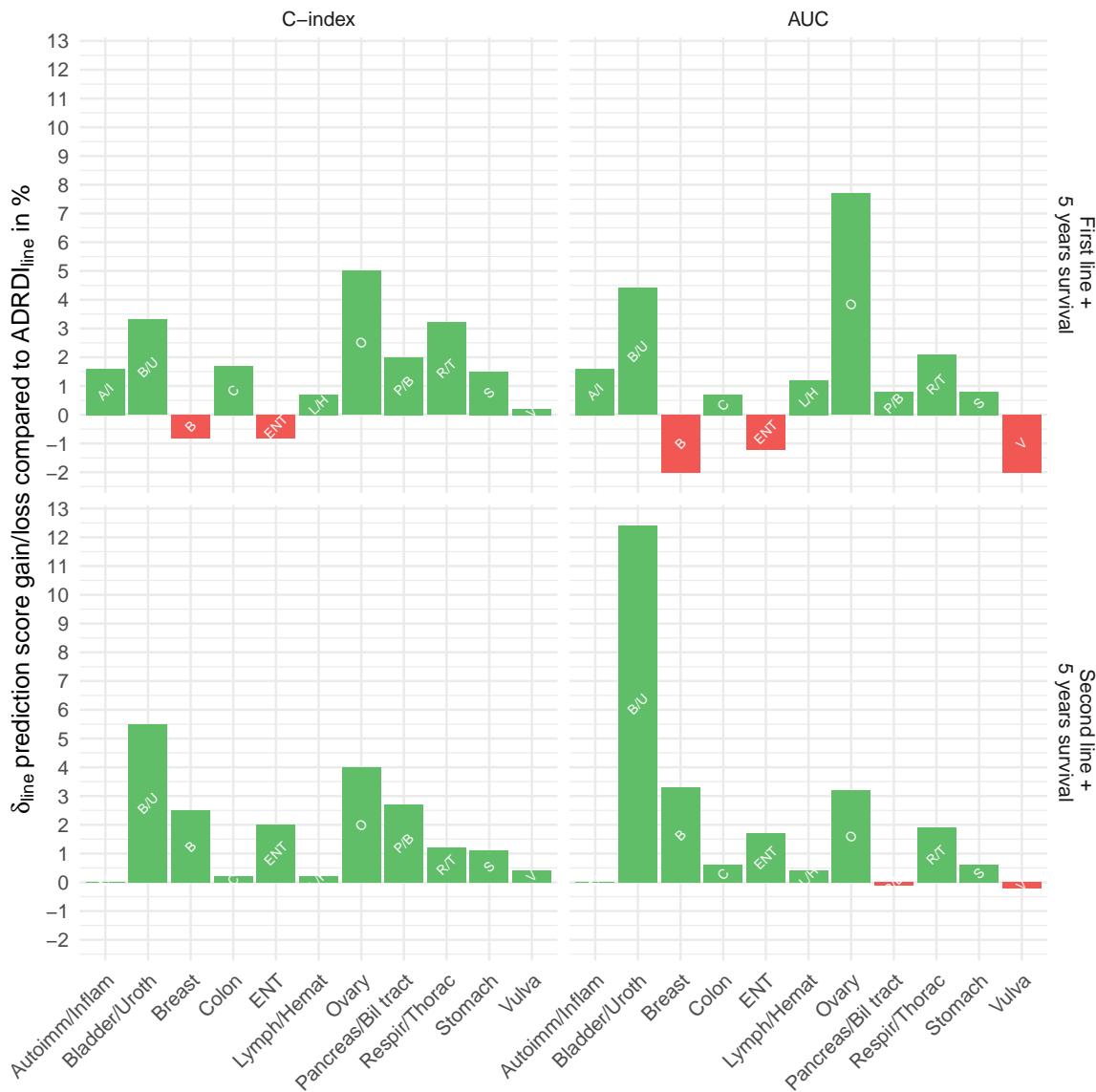


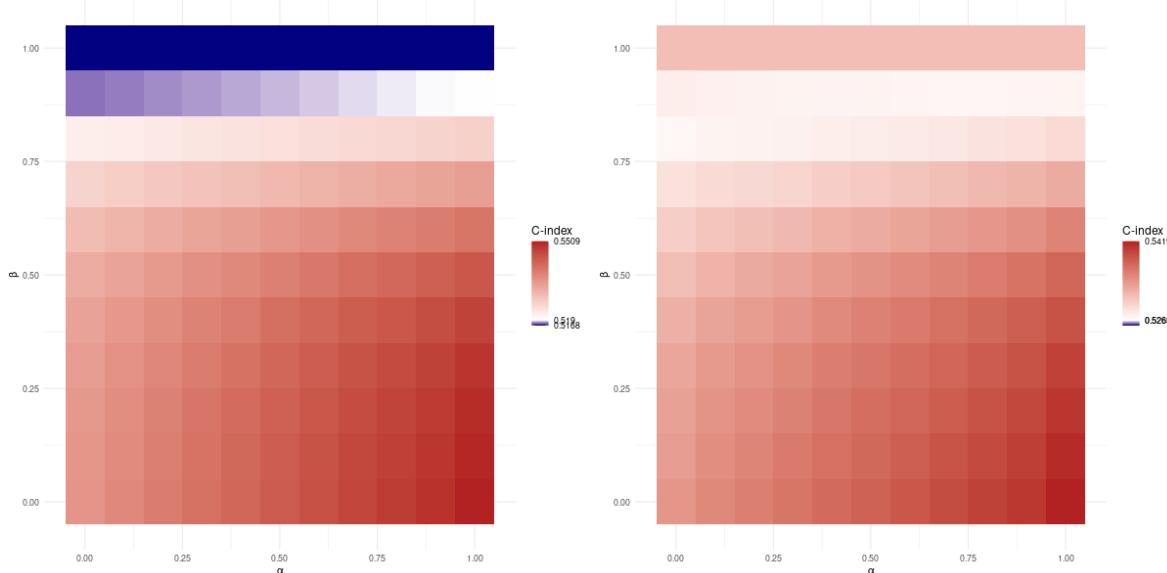
FIGURE 5.20 – Gains des scores de prédition de la survie globale à 5 ans des régressions de Cox (C-index) et logistique (AUC-ROC)

L'optimisation des poids de ProtoDrift avec la régression de Cox sur les 22 cohortes de l'HEGP a conduit à des améliorations notables : les scores de C-index ont augmenté dans 19 cohortes (87 %), avec des gains significatifs d'au moins 1 % observés dans 14 cohortes (63 %). Cette tendance est corroborée par les résultats de l'optimisation des poids avec la régression logistique, qui montrent également des résultats positifs dans 16 cohortes (72 %), avec des gains substantiels d'au moins 1 % dans 10 cohortes (45 %). Les améliorations du C-index sont plus répandues, tandis que l'ampleur des gains des scores AUC-ROC tend à être plus élevée. On note que les cohortes des cancers de l'ovaire et de la vessie, en première

et deuxième ligne de traitement, montrent des augmentations dépassant 3 % à la fois dans les scores de C-index et d'AUC-ROC.

5.8.4 Exploration de l'impact du temps par rapport à la dose sur la survie globale

La figure 5.21 illustre les prédictions de survie globale pour différentes combinaisons de poids de ProtoDrift associés au calendrier et au dosage des médicaments de chimiothérapie à travers les cohortes respiratoire et thoracique en première ligne à l'HEGP et au CHU de Bordeaux (*cf.* figure 5.9).



(a) *Heatmap* de la cohorte R/T en première ligne à l'HEGP (b) *Heatmap* de la cohorte R/T en première ligne au CHU de Bordeaux

FIGURE 5.21 – Gains de prédiction en C-index obtenus avec les modèles ProtoDrift par rapport au modèle BADRDI selon les valeurs de α (*i.e.*, $\frac{\omega_t}{\omega_t + \omega_d}$) et β (*i.e.*, $\frac{\omega_{\text{inter}}}{\omega_{\text{inter}} + \omega_{\text{intra}}}$) avec la cohorte respiratoire et thoracique en première ligne à l'HEGP (à gauche, figure 5.21a) et au CHU de Bordeaux (à droite, figure 5.21b).

Le maximum de gain de prédiction étant obtenu en bas droite cela implique (*cf.* figure 5.9) :

- que les déviations de calendrier au sein du cycle ont un impact plus significatif sur la survie globale que les déviations de dosage.
- que les écarts de calendrier et de dosage des administrations de médicaments anti-cancéreux ont plus d'impact sur la survie globale que la régularité de l'administration de l'ensemble du cycle.

5.9 Discussion et perspectives

Nous avons présenté ProtoDrift, une nouvelle métrique démontrant une fiabilité supérieure pour mesurer l'adhésion aux chimiothérapies par rapport à la dose-intensité relative (RDI) traditionnellement utilisée, notamment avec une meilleure prise en compte du calendrier. ProtoDrift montre non seulement des gains de performance, mais sa conception — associer des poids aux déviations de traitement — offre des perspectives de compréhension des impacts de ces déviations.

Cette discussion est divisé en trois sections. Premièrement, nous discutons des résultats d'application que nous venons d'exposer. Deuxièmement nous discutons de la méthode d'optimisation des poids de ProtoDrift et troisièmement nous discutons de sa conception (ordre anti-chronologique).

5.9.1 Interprétation des résultats d'application

Dans cette section, nous complétons et discutons de l'analyse des résultats d'application de la section 5.8. Nous commençons par émettre des critiques sur la méthode utilisée pour l'analyse comparative des trois méthodes (BADRDI, NP et PS-OP). Puis, nous interprétons les résultats obtenus pour les cohortes respiratoires et thoraciques en première ligne de traitement à l'HEGP et au CHU de Bordeaux. Enfin, nous étendons notre discussion à l'analyse des résultats pan-cancer.

5.9.1.1 Critique de la méthode d'analyse comparative

Dans cette section, nous formulons plusieurs remarques et critiques concernant la méthodologie de l'analyse comparative (*cf.* section 5.4), qui sont à prendre en compte dans l'interprétation des résultats.

L'un de nos trois critères dans l'analyse comparative est la significativité de la contribution de la variable explicative au modèle de régression (ADRDI pour BADRDI, $\delta_{\text{line}}(\alpha = \frac{1}{2}, \beta = \frac{1}{2})$ pour NP et $\underset{\alpha, \beta}{\text{argmax}} \text{PS}(\delta_{\text{line}}(\alpha, \beta))$ pour PS-OP). Or, au cours du chapitre 3, nous avons pu constater dans les sections sur l'analyse de survie (*cf.* sections 3.2.3 et 3.2.4) à travers les études de cas et les nombreux exemples de la littérature, que ce critère est rarement utilisé dans les études cliniques. Pour améliorer notre analyse comparative, nous devrions calculer les hasards ratios relatifs à chaque paire de quartiles, de manière analogue à ce qui est fait pour analyser le pouvoir discriminant avec le test du log-rank.

De plus, il existe une limitation dans l'utilisation des estimateurs de Kaplan-Meier et des tests de log-rank pour analyser le pouvoir discriminant de chaque méthode. Cette limitation est visible dans nos résultats à l'HEGP pour les cohortes de **cancer du sein en deuxième ligne, vessie et urothéliale en première ligne et respiratoires et thoraciques en deuxième ligne**. Pour ces cohortes, nous observons un gain en prédition avec le C-index, mais une perte en pouvoir discriminant avec les tests de log-rank. En réalité, NP discrimine mieux, mais la division en quartiles ne permet pas de le mettre en évidence. Cette simplification en quatre catégories masque les nuances et les améliorations de discrimination que NP pourrait apporter. Par conséquent, bien que NP améliore effectivement la discrimination,

cette amélioration n'est pas nécessairement visible lorsque les données sont divisées en quartiles pour les tests de log-rank.

5.9.1.2 Cohorte respiratoire et thoracique en première ligne

Dans cette section, nous détaillons les résultats obtenus pour la cohorte respiratoire et thoracique (R/T) en première ligne de traitement, en commentant l'analyse comparative des tableaux 5.15 et 5.17, les courbes de Kaplan-Meier des figures 5.14 et 5.16, et les *heatmaps* de la figure 5.21.

Les figures 5.14 (HEGP) et 5.16 (CHU de Bordeaux) montrent de gauche à droite les courbes de Kaplan-Meier obtenues à partir des quartiles de 1-ADRDI pour évaluer la méthode BADRDI, δ_{line} pour évaluer la méthode NP, et C-index- δ_{line} pour évaluer la méthode C-index-OP. On observe de gauche à droite un écartement progressif entre les courbes. Ces écarts progressifs indiquent un pouvoir discriminant supérieur de NP sur BADRDI, puis de C-index-OP sur NP (et par transitivité, de C-index-OP sur BADRDI). Ce résultat visuel est confirmé statistiquement avec les tableaux d'analyse comparative.

La supériorité des performances de NP sur BADRDI implique que la mesure ProtoDrift est plus adaptée pour mesurer l'adhésion thérapeutique des patients R/T en première ligne que la méthode standard BADRDI. Ceci suggère que la prise en compte du temps est importante pour cette localisation de cancer en première ligne. La supériorité des performances de C-index-OP sur NP quant à elle, montre que les pondérations relatives de dosage et de calendrier ont du sens, et que très probablement dans le cas des cancers R/T, les déviations de calendrier sont importantes.

Les *heatmaps* 5.21 obtenues dans les deux hôpitaux viennent en effet confirmer ce résultat. Leur analyse révèle que les prédictions supérieures surviennent souvent lorsque le poids associé au jour (ω_t) de l'administration dépasse celui de son dosage (ω_d), indiquant que les déviations de calendrier au sein du cycle ont un impact plus significatif sur la survie globale que les déviations de dosage.

Cependant, bien que les *heatmaps* des deux hôpitaux soient similaires, des différences notables émergent dans la manière dont ProtoDrift et BADRDI catégorisent les groupes de patients dans les analyses de Kaplan-Meier (cf. figures 5.14 et 5.16). À l'HEGP, l'intuition que nous avions est effectivement observée : une meilleure adhésion aux métriques de traitement corrèle avec une amélioration de la survie. À l'inverse, au CHU de Bordeaux, le quartile le moins adhérent montre étonnamment des taux de survie plus élevés. Ce phénomène peut refléter les débats en cours sur la gestion de la chimiothérapie chez les patients âgés, en particulier ceux atteints de cancer du poumon non à petites cellules (NSCLC), où des dosages réduits et des administrations retardées peuvent paradoxalement bénéficier à la survie (cf. section 1.2.6) [79]. Ces résultats soulignent le potentiel de ProtoDrift pour révéler des impacts nuancés du traitement, mais nécessitent des investigations supplémentaires. Les analyses de nos résultats manquent en effet d'une interprétation médicale experte.

Après avoir analysé les résultats obtenus sur une localisation de cancer spécifique, nous élargissons notre analyse à une analyse pan-cancer.

5.9.1.3 Analyse pan-cancer

Dans cette section, on se concentre sur les résultats pan-cancer exposés dans les figures 5.18, 5.19 et 5.20.

Il n'est pas possible de tirer une conclusion globale sur l'adhésion aux chimiothérapies. Bien que les résultats d'évaluation (*cf.* section 5.8.2) suggèrent que la mesure ProtoDrift, même dans sa forme non optimisée, est globalement plus performante pour mesurer l'adhésion que la méthode standard BADRDI, ce résultat n'est pas unanime. On observe cette variabilité dans la figure 5.19 où on relève des pertes de performance pour les cohortes ORL en première ligne de l'HEGP, myélome en première et deuxième ligne, neurologie en première ligne et vessie et urothéliale du CHU de Bordeaux. Cette variabilité souligne les spécificités de chaque localisation, et de chaque traitement. Par exemple à l'HEGP, même si des pertes sont observées en première ligne pour la localisation ORL, on constate qu'en deuxième ligne, le résultat est complètement inversé avec des gains de performance pour nos trois critères. Cette variabilité souligne l'importance d'adapter la mesure, ce qui renforce l'intérêt d'optimiser les poids de ProtoDrift. La figure 5.20 confirme la pertinence d'optimiser ces poids, avec gains de performance obtenus pour la majorité des cohortes en C-index et en AUC-ROC.

La spécificité des localisations de cancer et de leur traitement se manifeste également par la diversité des profils des heatmaps, accessibles sur le site <https://files.inria.fr/protodrift-surv/>. Les régions des *heatmaps* où l'on observe des gains de performance varient en fonction des cohortes étudiées. La figure 5.22 synthétise cette variabilité en présentant une série de graphiques en facettes, illustrant les régions des *heatmaps* avec un gain de performance pour chaque cohorte ayant obtenu un gain de prédiction (en C-index ou en AUC-ROC) supérieur à 1%.

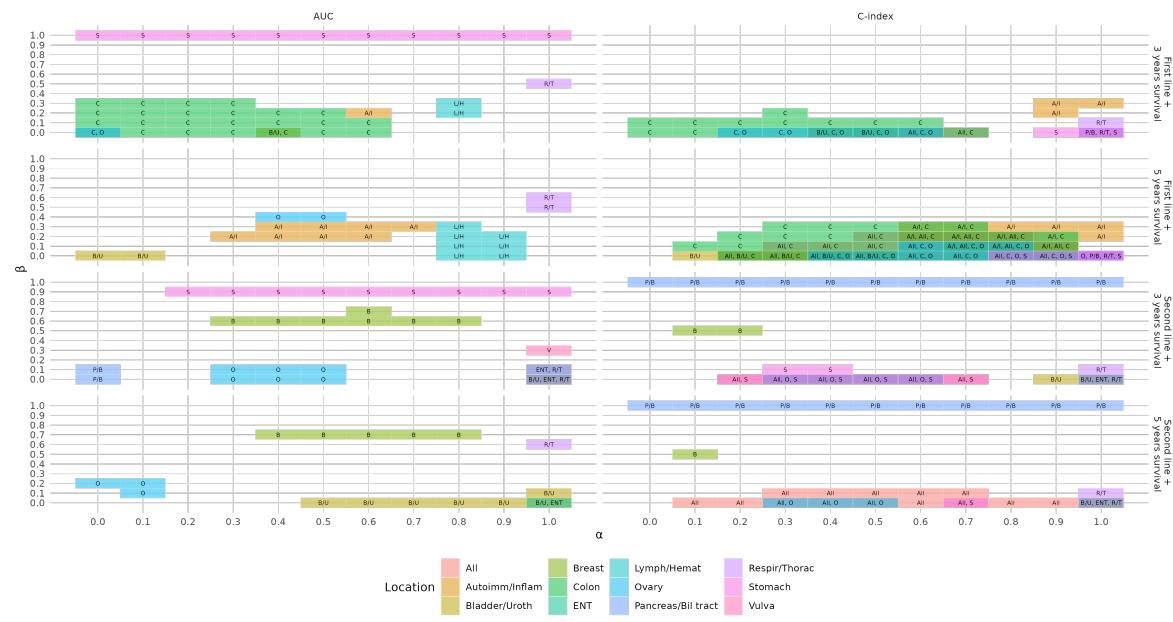


FIGURE 5.22 – Graphiques en facettes illustrant les régions des *heatmaps* avec un gain maximal pour les cohortes ayant obtenu un gain de prédiction supérieur à 1%. À gauche, les graphiques présentent les régions avec un gain maximal en AUC-ROC. À droite, les graphiques présentent les régions avec un gain maximal en C-index. Chaque rangée de graphiques représente une ligne de traitement et un temps de survie spécifiques. Les couleurs représentent les différentes localisations de cancer.

On constate que certaines cohortes, comme celles des cancers de la vessie et urothelial en deuxième ligne, ou ORL en deuxième ligne, présentent des profils similaires à la cohorte respiratoire en première ligne, avec un maximum en bas à droite de la *heatmap*, observé à la fois en AUC-ROC et en C-index. Pour ces cohortes, on peut donc conclure que les déviations de calendrier au sein du cycle ont un impact plus significatif sur la survie globale que les déviations de dosage.

Pour d'autres cohortes, comme celles du cancer du colon en première ligne et des ovaires en première ligne, une tendance inverse est observée, avec des régions maximales plus étendues et décalées vers la gauche. Cela indique que les déviations de dosage ont une importance relative plus grande que les déviations de calendrier au sein du cycle, et que ces déviations au sein du cycle sont plus cruciales que la régularité des cycles.

À l'inverse, pour certaines cohortes, la régularité des cycles semble avoir de l'importance, avec des maxima observés en haut de la *heatmap*. Ce résultat est visible pour la cohorte des cancers du pancréas et des voies biliaires en deuxième ligne avec le C-index, ainsi que pour le cancer de l'estomac en première et deuxième ligne avec l'AUC-ROC. Toutefois, en consultant les *heatmaps* de ces cohortes sur <https://files.inria.fr/protodrift-surv/years-survival-68> et <https://files.inria.fr/protodrift-surv/years-survival-80>, on constate que même si un maximum est observé en haut, ces cohortes possèdent deux régions avec des gains de prédiction élevés. La figure 5.22 offre une vue d'ensemble globale, mais elle agrège les informations discernables sur les *heatmaps* individuelles.

5.9.2 Optimisation des poids de ProtoDrift

Dans cette section, nous discutons de la méthode utilisée pour optimiser les poids de ProtoDrift.

5.9.2.1 Le poids médicaments

Nous avons utilisé l'ADRDI (*All Drugs Relative Dose-Intensity*) pour comparer notre mesure ProtoDrift à la dose-intensité relative (RDI), la mesure standard d'adhésion au protocole de chimiothérapie. L'ADRDI est une moyenne de la RDI de tous les médicaments administrés au cours d'une ligne de chimiothérapie (*cf. sections 1.3.2 et 5.1.3.2*). Cependant, l'ADRDI est rarement utilisée dans les études cliniques similaires à celles citées dans le chapitre 1 (*cf. section 1.3.2*) et le chapitre 3 (*cf. section 3.2.1*). Ces études calculent généralement une RDI par médicament administré, pour une localisation de cancer et une ligne de traitement spécifique. Pour aligner notre étude avec ces pratiques, nous pourrions ajouter une stratification supplémentaire par médicament à notre stratification par localisation de cancer et par ligne de traitement. Cela impliquerait des profils de *heatmap* (α, β) par cancer, par ligne et par médicament. Toutefois, la taille réduite de nos cohortes d'étude empêche cette stratification supplémentaire. C'est pour cette même raison que nous n'avons pas optimisé les poids des médicaments ω_m dans ProtoDrift, comme expliqué en section 5.3.2 sur les objectifs d'optimisation.

L'optimisation des poids des médicaments ω_m pourrait fournir des informations précieuses sur les choix relatifs aux réductions et décalages d'administrations spécifiques aux médicaments. Dans ce travail, nous avons particulièrement souligné l'importance des variations de poids. Les résultats et discussions récentes ont montré que l'exploration des poids à l'aide des *heatmaps* révèle effectivement des informations cruciales sur les déviations. Avec seulement quatre poids (deux paramètres libres α et β), la quantité de résultats générés rend déjà l'analyse pan-cancer très riche et dense. Ajouter les poids des N médicaments ω_m introduirait $N - 1$ paramètres libres supplémentaires, ce qui représente une perspective complexe mais excitante.

5.9.2.2 Modèles de survie

Dans cette section nous critiquons et discutons des modèles de survie ajustés et entraînés dans l'optimisation (*cf. section 5.3.3.3*).

Nous avons utilisé des régressions logistiques et de Cox pour les entraîner et obtenir des scores de prédictions sur la survie globale des patients avec différentes combinaisons de valeurs α, β de la variable explicative δ_{line} . Ceci est comparable aux méthodes de construction d'un nomogramme décrites dans la section 3.2.4 du chapitre 3 de l'état de l'art.

Régression logistique et modèles de temps

Dans les constructions de nomogramme, il est fréquent d'entraîner une régression logistique sur un modèle de temps continu comme nous l'avons fait. Néanmoins, l'état de l'art sur l'utilisation de régression logistique pour modéliser la survie (*cf. sections 3.2.2 et 3.2.3*) recommande d'utiliser un modèle de temps discret. De plus, cet état de l'art a aussi mis en

évidence la pertinence d'utiliser les régressions logistiques en survie dans le cas où la survenue de l'événement d'intérêt n'est pas clairement identifiée, mais qu'on sait qu'elle est survenue dans un intervalle de temps. Or dans notre cas, comme nous l'avons déjà remarqué, les dates de décès sont identifiées et proviennent de l'Insee, une source de qualité (cf. 1.4.8). Bien qu'on l'observe fréquemment dans la littérature, l'utilisation d'une régression logistique sur un modèle de temps continu ne semble pas pertinente, et encore moins dans notre cas où les temps de survenue d'événement sont clairement identifiés. En revanche, elle serait pertinente dans le cas où on prendrait, par exemple, la survenue d'une toxicité irréversible comme événement d'intérêt, et où on formaterait les données de manières à la réaliser avec un modèle de temps discret (cf. section 3.8).

Modèles joints

Dans la section 3.2.1 du chapitre 3, nous avons évoqué d'autres modèles paramétriques qui supposent une distribution spécifique de la fonction de risque $h(t)$. Ces modèles ne semblent pas adaptés à notre contexte. En revanche, il existe des modèles joints composés de deux sous-modèles : un modèle longitudinal, souvent spécifié avec un modèle linéaire mixte (*linear mixed model*), et un modèle de survie [96]. Ces modèles sont particulièrement utiles dans les situations où il est nécessaire de modéliser la relation entre une trajectoire de biomarqueurs mesurés de manière répétée et le temps jusqu'à un événement d'intérêt. En combinant ces deux sous-modèles, le modèle joint prend en compte à la fois l'évolution des biomarqueurs dans le temps et leur impact sur le risque de survenue de l'événement d'intérêt.

Dans notre cas, il serait pertinent de modéliser l'évolution des dissimilarités de cycle δ_{cycle} et de les associer avec la survenue d'un événement d'intérêt comme le décès ou une toxicité irréversible. Nous pourrions également utiliser l'ADRFI comme ligne de base, en modélisant son évolution à chaque cycle avec un modèle linéaire mixte. Cela permettrait de mieux comprendre comment les modifications du traitement au fil du temps influencent les résultats cliniques des patients.

Modèles de survie par apprentissage automatique Les principaux algorithmes utilisés en apprentissage automatique appliqués à la survie sont les forêts aléatoires de survie (*Random Survival Forest*) et les amplifications de gradient (*Survival Gradient Boosting*). Les deux algorithmes ont pour objectif final d'estimer la fonction de survie $S(t)$ ou la fonction de risque cumulé $H(t)$ et non la fonction de risque instantané comme le fait le modèle de Cox.

Les forêts aléatoires de survie construisent plusieurs arbres de décision sur des échantillons bootstrap des données et agrègent les résultats pour obtenir une estimation de la fonction de survie. Cette méthode est particulièrement adaptée lorsque le nombre de variables explicatives est élevé. On a pu constater dans les constructions de nomogramme (cf. section 3.2.4) qu'elles étaient principalement utilisées lors de la sélection de variables, mais que le modèle entraîné était un modèle de Cox ou une régression logistique. Ces modèles pourraient être adaptés pour l'optimisation des poids de médicaments ω_m de ProtoDrift (cf. section 5.9.2.1). Les amplifications de gradient pour la survie adaptent la technique de descente de gradient pour minimiser une fonction de perte spécifique aux données de survie, en construisant successivement des modèles "faibles" (comme des arbres de décision) pour améliorer l'estimation de la survie. Cette méthode est particulièrement adaptée lorsque les relations entre

les variables explicatives et le temps de survie sont complexes.

Toutefois, l'interprétabilité des modèles de forêts aléatoires de survie et des amplifications de gradient peut être complexe en raison de la nature non linéaire et des interactions entre les variables dans ces modèles.

5.9.2.3 Méthode d'évaluation de ProtoDrift

Évaluation de ProtoDrift sur deux jeux de données indépendants :

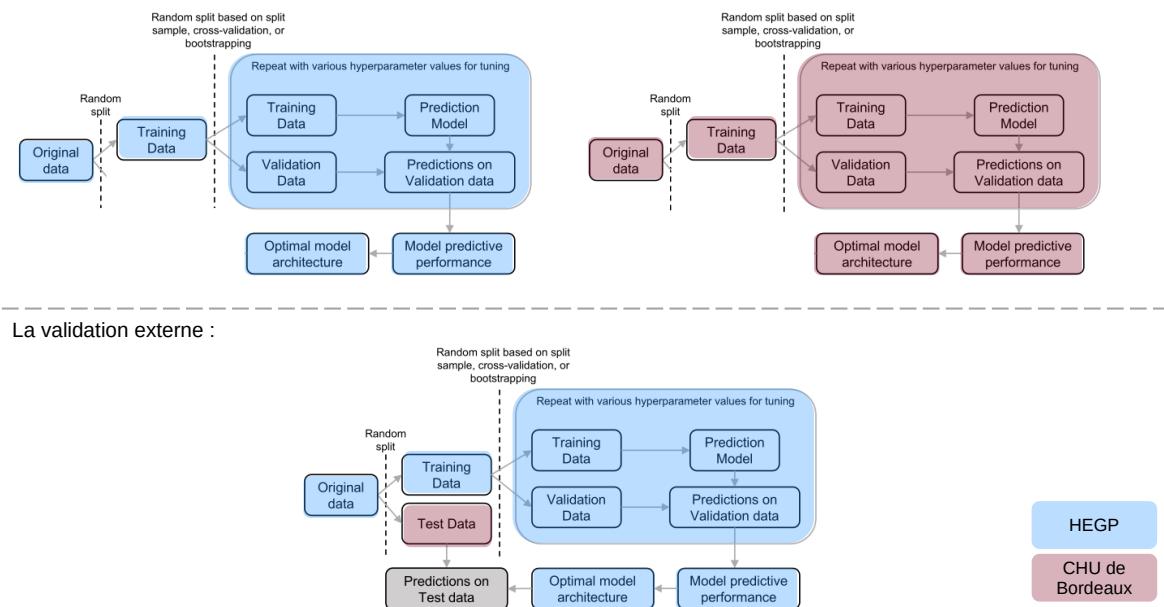


FIGURE 5.23 – Comparaison entre notre méthode d'évaluation de ProtoDrift et la méthode de validation externe recommandée dans la littérature. Figure adaptée de l'article de SURESH, SEVERN et GHOSH [195].

Pour conclure cette discussion sur l'optimisation, nous revenons sur la méthode que nous avons utilisée pour évaluer la mesure ProtoDrift (*cf.* section 5.7.3). Dans la section 3.2.4, nous avons présenté la méthode de validation recommandée dans la littérature pour évaluer les modèles prédictifs en général, et les nomogrammes en particulier.

La figure 5.23 schématisé une comparaison entre notre méthode d'évaluation et la méthode de validation recommandée pour les modèles prédictifs. En suivant cette méthode, nous aurions dû calculer les métriques de prédiction des modèles optimaux (C-index-OP) obtenus à l'HEGP sur des jeux de données test provenant du CHU de Bordeaux. Cependant, nous avons choisi d'entraîner indépendamment des modèles de Cox et logistique avec δ_{line} et l'ADRDI sur deux jeux de données distincts, puis de comparer les gains de prédiction obtenus.

Cette stratégie d'évaluation s'explique par l'objectif spécifique de ProtoDrift, qui est de mesurer l'adhésion thérapeutique plutôt que de prédire la survie. Nous partons du principe qu'il peut exister une grande variabilité de la mesure ProtoDrift entre deux cohortes de patients atteints du même cancer mais traités dans des hôpitaux différents. Cette variabilité

pourrait être due à la fois aux différences dans les populations de patients et aux variations dans les pratiques de soins.

Notre objectif n'est donc pas de créer un modèle de prédiction général, mais d'évaluer si ProtoDrift peut capturer cette variabilité de manière indirecte. Cette approche nous permet d'explorer l'impact de l'adhésion thérapeutique sur les résultats cliniques de manière plus précise et adaptée à chaque contexte spécifique.

Après avoir discuté des différents aspects de la méthode d'optimisation des poids de ProtoDrift, nous abordons sa conception dans la dernière section.

5.9.3 Conception de la mesure ProtoDrift

Dans cette section, nous discutons de la conception de la mesure ProtoDrift, de son objectif et des choix que nous avons faits pour définir les différentes dissimilarités qui représentent les déviations possibles au traitement.

5.9.3.1 Ajout de poids temporels

Bien que nous ayons défini certains poids pour représenter l'importance relative des déviations au traitement, il se peut que d'autres poids pertinents nous aient échappé. Par exemple, les impacts des déviations en début de ligne peuvent différer de ceux en fin de ligne. Il est possible que le nombre de cycles suivis par un patient influence les décisions de décalage ou de réduction de dose. Ainsi, on pourrait envisager d'introduire un poids associé au numéro du cycle suivi. De même, un poids pourrait être associé au numéro de la ligne de traitement.

5.9.3.2 Modélisation continue du temps d'exposition au médicament

Notre approche guidée par les données (*data-driven*) représente les déviations aux protocoles de manière discrète. Plutôt que de représenter les administrations médicamenteuses comme des prises ponctuelles, on pourrait modéliser le temps d'exposition médicamenteux par une fonction continue. La méthode d'exposition cumulative pondérée (*Weighted Cumulative Exposure*) permet de modéliser ce temps d'exposition [109]. La dissimilarité de la molécule δ_m pourrait alors être calculée en fonction du recouvrement entre la fonction représentant l'exposition théorique et celle du cycle suivi.

5.9.3.3 Algorithme d'alignement

Nous avons défini un algorithme qui aligne les paires d'administrations les plus proches pour calculer la dissimilarité de médicament δ_m (cf. algorithme 5.2). Cependant, il existe des algorithmes de déformation temporelle (*Dynamic Time Warping*) qui permettent d'aligner deux séries temporelles. Nous pourrions utiliser ces algorithmes pour aligner les paires d'administrations et calculer la dissimilarité de médicament δ_m de manière plus précise.

5.9.3.4 De la mesure de l'adhésion à la mesure de similarité entre lignes suivies

ProtoDrift a été développée pour mesurer l'adhésion au protocole théorique de chimio-

thérapie. Ainsi, les dissimilarités ont souvent été définies par rapport au cycle théorique. Par exemple, les dissimilarités de calendrier, c'est-à-dire les dissimilarités de jour d'administration δ_t (*cf.* formule 5.12) et inter-cycle δ_{inter} (*cf.* formule 5.7), ont été définies de manière asymétrique. On pourrait envisager une nouvelle mesure ("LineDist"), qui compareraient les lignes de traitement suivies par les patients, offrant ainsi une nouvelle perspective sur la similarité entre les lignes de traitement.

5.10 Bilan

Ce chapitre a présenté ProtoDrift, une nouvelle méthode pour mesurer l'adhésion aux traitements de chimiothérapie, en se basant sur une approche de dissimilarité pondérée. Les objectifs principaux étaient de surmonter les limitations de la mesure traditionnellement utilisée, la dose-intensité relative (RDI) et de proposer un outil capable de capturer les déviations temporelles et de dosage dans les traitements de chimiothérapie.

5.10.1 Résumé des contributions

- **Conception de ProtoDrift :** Nous avons défini une nouvelle métrique, ProtoDrift, qui intègre des dissimilarités à différents niveaux du traitement de chimiothérapie. Cette métrique permet de quantifier de manière plus précise les écarts par rapport aux protocoles de traitement prévus. (*cf.* section 5.2 et projet gitlab **ProtoDrift** disponible sur <https://gitlab.inria.fr/arogier/protodrift>)
- **Optimisation des poids :** Nous avons mis en place une méthode d'optimisation pour ajuster les poids des différentes dissimilarités, en fonction de leur impact relatif sur la survie des patients. Cette optimisation a été réalisée en utilisant des modèles de régression logistique et de Cox, et évaluée par une approche de rééchantillonnage *bootstrap*. (*cf.* section 5.3 et projet gitlab **ProtoDrift-Surv** disponible sur <https://gitlab.inria.fr/arogier/protodrift-surv>)
- **Application et évaluation sur deux jeux de données indépendants :** ProtoDrift a été appliquée et évaluée sur des données provenant de l'HEGP et du CHU de Bordeaux. Les résultats ont montré une amélioration des performances prédictives par rapport à la méthode RDI traditionnellement utilisée, soulignant la capacité de ProtoDrift à mieux capturer les variations et à fournir une mesure plus fine de l'adhésion aux traitements. (*cf.* sections 5.4, 5.7 et 5.8)
- **Résultats d'analyse pan-cancer :** L'application de ProtoDrift sur les données de l'HEGP et du CHU de Bordeaux a généré une grande quantité de résultats pan-cancer, offrant de nouvelles perspectives pour la recherche clinique en oncologie. Ces résultats montrent des variations dans l'adhésion aux chimiothérapies selon les localisations de cancer et soulignent l'importance d'une optimisation des poids de ProtoDrift pour chaque contexte spécifique. (*cf.* section 5.8 et 5.9.1 et site web <https://files.inria.fr/protodrift-surv/>)

5.10.2 Implications pour la recherche clinique

Les résultats obtenus avec ProtoDrift sont prometteurs pour la recherche en oncologie. En fournissant une évaluation plus nuancée des ajustements de traitement, ProtoDrift peut aider à mieux comprendre les impacts de ces ajustements sur l'évolution clinique des patients. Les résultats pan-cancer permettent d'identifier des tendances et des différences

selon les localisations de cancers et les lignes de traitement, ce qui est précieux pour orienter les futures études cliniques.

Un des avantages majeurs de ProtoDrift réside dans son interprétabilité. En utilisant seulement quatre poids liés aux décisions cliniques pendant le traitement, il reste compréhensible contrairement à certains modèles d'apprentissage machine à haute paramétrisation. Cette clarté est un atout dans un outil de support à la recherche clinique.

Enfin, la faisabilité et l'application large de ProtoDrift sont soulignées par sa mise en œuvre réussie dans deux hôpitaux distincts. Les données nécessaires pour calculer ProtoDrift sont facilement disponibles à partir du logiciel Chimio®, utilisé par la grande majorité des hôpitaux autorisés à administrer la chimiothérapie en France, ce qui souligne son potentiel de portée.

5.10.3 Défis et perspectives

- **Optimisation des poids ω_m** : Bien que nous n'ayons pas optimisé les poids des médicaments spécifiques ω_m dans ce travail, cette optimisation pourrait être réalisée dans des études futures si des cohortes suffisamment grandes sont disponibles, et avec des algorithmes d'apprentissage automatique. (cf. sections 5.3.2, 5.9.2.1, et 5.9.2.2.3)
- **Approfondissement de l'analyse pan-cancer avec les yeux d'experts** : Les résultats obtenus doivent être approfondis avec l'aide d'oncologues pour interpréter les variations observées et valider les tendances identifiées. Cette collaboration enrichira l'analyse actuelle et permettra de mieux comprendre l'impact des ajustements de traitement spécifiques à différents contextes cliniques. (cf. section 5.9.1 et site web <https://files.inria.fr/protodrift-surv/>)
- **Longitudinalité des dissimilarités** : Pour étudier davantage l'évolution et la longitudinalité des dissimilarités de ProtoDrift, il serait pertinent d'explorer des méthodes telles que l'introduction de poids temporels, le *Dynamic Time Warping* (DTW), l'exposition cumulative pondérée (WCE) et les modèles joints. (cf. sections 5.9.2.2.2 et 5.9.3.1)

5.10.4 Conclusion

ProtoDrift représente une avancée significative dans l'évaluation de l'adhésion aux traitements de chimiothérapie. En capturant de manière détaillée les ajustements de traitement, cette méthode offre une meilleure compréhension de leur impact sur la survie des patients. Les travaux futurs devraient se concentrer sur l'optimisation de poids médicaments, une analyse experte des résultats obtenus, et l'exploration de la longitudinalité des dissimilarités. Avec ProtoDrift, nous espérons contribuer à une meilleure compréhension des traitements de chimiothérapie et à une amélioration de l'interprétation de résultats de la recherche clinique en oncologie.

Quatrième partie

Conclusion et perspectives

CONCLUSION ET PERSPECTIVES

Le sujet de cette thèse était d'extraire et prédire les réponses aux chimiothérapies à partir des données hospitalières.

Au cours du chapitre 1, nous avons mis en évidence les défis latents de ce sujet. Les chimiothérapies sont complexes en raison de leur organisation rigoureuse en cycles et lignes de traitement, et elles produisent des réponses variées chez les patients. La gestion de leur efficacité et de leur tolérance nécessite une coopération étroite entre patients et médecins, avec une surveillance et des adaptations régulières. En parallèle, la réutilisation des Dosiers Patients Informatisés (DPI) pour la recherche est rendue difficile par l'hétérogénéité et la variabilité des données. Pour exploiter pleinement les DPI, des techniques sophistiquées d'extraction, de représentation et d'intégration des informations sont nécessaires.

L'état de l'art nous a fourni des pistes pour relever ces défis. Dans le chapitre 2, nous avons montré que les ontologies ont un intérêt pour représenter les connaissances biomédicales en structurant, intégrant et interconnectant des ressources disponibles et déjà formalisées. Ces capacités permettent de surmonter les limites des entrepôts de données traditionnels et ouvrent perspectives pour l'analyse et l'exploration des données médicales. Dans le chapitre 3, nous avons présenté des techniques d'extraction de connaissances à partir de sources hétérogènes. Ces techniques incluent des outils de traitement automatique des langues pour l'extraction de connaissances à partir de comptes rendus cliniques, mais aussi des mesures qui détectent du signal dans les grands volumes de données. Ce chapitre a également montré comment prédire des événements à partir des connaissances extraites à l'aide de modèles de survie.

En nous appuyant sur les méthodes présentées dans l'état de l'art, nous avons tenté de répondre aux défis soulevés dans le chapitre 1 à travers deux contributions principales.

La première contribution concerne la représentation des chimiothérapies et de leurs réponses avec les ontologies ChemoOnto et OntoTox, présentées dans le chapitre 4. ChemoOnto se concentre sur le déroulement des chimiothérapies, tandis qu'OntoTox se focalise sur les toxicités et leurs sévérités. Les deux ontologies sont liées à des ressources de référence et constituent des bases de connaissances précieuses pour l'analyse des réponses aux chimiothérapies.

Cependant, il reste des améliorations à apporter aux techniques d'extraction des connaissances pour renforcer la fiabilité des données intégrées aux deux ontologies. De plus, ces bases de connaissances, bien que précieuses, ne peuvent pas être partagées en raison du caractère personnel des données des patients. Ainsi, seule la représentation des protocoles théoriques de ChemoOnto a pu être partagée, donnant lieu à une base de connaissances ouverte et accessible, [ChemoKG \(<http://chemokg.paris.inria.fr/>\)](http://chemokg.paris.inria.fr/). Des algorithmes de plongement de graphes sur ChemoKG ont permis de regrouper les protocoles de chimiothérapie de manière cohérente, selon les classes de médicaments ATC et les localisations de cancer. De la même manière, il serait intéressant d'explorer ChemoOnto instantanée avec les lignes de traitement des patients et d'analyser un clustering des traitements suivis. Ces résultats pourraient être liés aux résultats d'adhésion mesurés avec ProtoDrift,

notre deuxième contribution.

La deuxième contribution est la conception de ProtoDrift, une mesure de l'adhésion aux chimiothérapies. ProtoDrift répond aux limitations de la dose-intensité relative (RDI) en quantifiant les écarts par rapport aux protocoles de traitement à travers des dissimilarités pondérées. Elle s'est avérée plus fiable que la RDI pour évaluer l'adhésion aux chimiothérapies, notamment en intégrant mieux les aspects calendaires. ProtoDrift améliore non seulement les performances prédictives, mais permet aussi de mieux comprendre les impacts des écarts de traitement grâce à sa conception qui attribue des poids spécifiques à ces écarts.

L'application de ProtoDrift a généré une grande quantité de résultats pan-cancer, montrant l'impact de l'adhésion aux chimiothérapies sur la survie globale. Les résultats obtenus sur chaque cohorte étudiée sont disponibles sur <https://files.inria.fr/protodrift-surv/>. Ces résultats nécessitent une analyse médicale experte. Avec ProtoDrift, nous espérons faciliter la compréhension des réponses aux chimiothérapies et l'interprétation des résultats de recherche clinique en oncologie. Dans le chapitre 1, nous avons mentionné certains désaccords entre oncologues sur les ajustements de protocoles (*cf. section 1.2.6*). Il serait intéressant de mener une étude avec ProtoDrift appliquée à une population spécifique pour analyser de tels désaccords. Par exemple, on pourrait appliquer ProtoDrift sur les données d'une cohorte de patients âgés atteints de cancer du poumon suivant des protocoles avec du carboplatine, puis vérifier si les réductions de doses ont un impact sur la survenue de toxicités. Toutefois, comme indiqué dans la section 1.5.4 du chapitre 1, les résultats de ProtoDrift peuvent compléter les résultats d'essais cliniques ou générer des hypothèses, mais celles-ci nécessitent des essais cliniques randomisés pour être validées.

Les résultats de cette thèse ouvrent de nombreuses perspectives, nécessitant des collaborations étroites entre informaticiens et cliniciens. Les outils développés, ChemoOnto, OntoTox et ProtoDrift, offrent des bases solides pour des études futures, et soulignent les bénéfices de l'interdisciplinarité pour enrichir notre compréhension des réponses aux chimiothérapies.

Cinquième partie

Annexes

ANNEXES

CHAPITRE 5

method	Line	Cox regression fit results		C-index	Number of significant survival time differences between quartile pairs
		θ_1	p-value		
Autoimm/Inflam					
OP	1	2.998	0.014	0.68	4
NP	1	3.793	0.029	0.677	3
BADRDI	1	-0.063	0.46	0.664	1
OP	2	2.94	0.302	0.62	0
NP	2	1.799	0.365	0.62	0
BADRDI	2	-1.101	0.346	0.62	0
Colon					
OP	1	1.297	0.029	0.521	3
NP	1	1.682	0.044	0.518	1
BADRDI	1	-0.263	0.304	0.504	0
OP	2	0.44	0.305	0.493	0
NP	2	0.102	0.447	0.489	2
BADRDI	2	0.199	0.408	0.491	0
Stomach					
OP	1	1.157	0.159	0.533	0
NP	1	0.697	0.387	0.521	0
BADRDI	1	0.144	0.44	0.518	0
OP	2	1.243	0.268	0.511	0
NP	2	1.558	0.31	0.506	0
BADRDI	2	-0.126	0.425	0.5	0
Lymph/Hemat					
OP	1	1.351	0.101	0.588	3
NP	1	1.647	0.15	0.586	2
BADRDI	1	-0.515	0.256	0.581	1
OP	2	-1.358	0.234	0.522	1
NP	2	-1.323	0.297	0.521	0
BADRDI	2	0.477	0.36	0.52	0
ENT					
OP	1	0.384	0.412	0.519	0
NP	1	0.203	0.422	0.518	0
BADRDI	1	-0.526	0.129	0.527	3
OP	2	1.099	0.007	0.538	3
NP	2	1.746	0.02	0.53	3
BADRDI	2	-0.659	0.106	0.518	2
Ovary					
OP	1	1.716	0.019	0.535	4
NP	1	2.759	0.038	0.524	3
BADRDI	1	-0.441	0.357	0.485	0
OP	2	1.634	0.057	0.544	4
NP	2	2.192	0.121	0.525	2
BADRDI	2	-0.73	0.276	0.504	2
Pancreas/Bil tract					
OP	1	0.931	0.051	0.561	0
NP	1	0.84	0.212	0.548	3
BADRDI	1	-0.166	0.408	0.541	0
OP	2	-1.7	0.077	0.554	3
NP	2	0.748	0.32	0.529	0
BADRDI	2	-0.31	0.346	0.527	1
Respir/Thorac					

OP	1	1.101	0	0.551	6
NP	1	1.442	0.006	0.538	4
BADRDI	1	-0.283	0.234	0.519	0
OP	2	0.713	0.072	0.527	2
NP	2	0.947	0.149	0.523	1
BADRDI	2	-0.255	0.348	0.515	2
Breast					
OP	1	5.032	0	0.655	6
NP	1	4.581	0	0.653	6
BADRDI	1	-2.567	0	0.663	5
OP	2	3.942	0	0.595	5
NP	2	3.92	0	0.594	4
BADRDI	2	-1.806	0.003	0.57	5
Bladder/Uroth					
OP	1	1.222	0.01	0.552	4
NP	1	1.167	0.173	0.53	2
BADRDI	1	-0.17	0.407	0.519	3
OP	2	1.982	0.047	0.553	4
NP	2	2.881	0.078	0.536	5
BADRDI	2	-0.375	0.371	0.498	2
Vulva					
OP	1	0.732	0.166	0.554	0
NP	1	1.153	0.231	0.55	2
BADRDI	1	-0.818	0.153	0.552	2
OP	2	0.397	0.38	0.509	0
NP	2	0.547	0.415	0.504	0
BADRDI	2	0.22	0.425	0.505	0

TABLE 5.5 – Prédiction des ajustements du modèle de Cox et détails gain/perte du pouvoir discriminant. Table de l'HEGP.

method	Line	Cox regression fit results		C-index	Survival time differences Number of significant survival time differences between quartile pairs
		θ_1	p-value		
Autoimm/Inflam					
NP	1	-0.487	0.349	0.766	0
BADRDI	1	0.311	0.228	0.767	0
NP	2	-1.86	0.084	0.643	0
BADRDI	2	0.649	0.158	0.641	0
Colon					
NP	1	-2.698	0.015	0.569	2
BADRDI	1	0.538	0.222	0.54	0
NP	2	-3.575	0.005	0.626	4
BADRDI	2	1.169	0.064	0.596	3
Melanoma					
NP	1	3.257	0	0.565	5
BADRDI	1	-0.254	0.391	0.506	0
NP	2	4.067	0.001	0.559	3
BADRDI	2	-0.455	0.355	0.513	0
Liver					
NP	1	-1.754	0.053	0.534	1
BADRDI	1	0.715	0.043	0.533	3
NP	2	-1.721	0.226	0.526	0
BADRDI	2	0.862	0.187	0.531	0
Myeloma					
NP	1	-3.468	0.002	0.595	3
BADRDI	1	1.902	0.001	0.606	4
NP	2	-2.233	0.149	0.524	0
BADRDI	2	1.553	0.039	0.542	3
Lymph/Hemat					
NP	1	0.998	0.003	0.621	5
BADRDI	1	-0.334	0.033	0.617	4
NP	2	0.797	0.116	0.578	4
BADRDI	2	-0.222	0.277	0.574	3
Neurology					
NP	1	-1.935	0.231	0.53	0
BADRDI	1	0.864	0.139	0.536	1
NP	2	-3.066	0.151	0.49	1
BADRDI	2	1.6	0.153	0.491	0
ENT					
NP	1	-7.376	0	0.582	5
BADRDI	1	-0.382	0.361	0.497	0
NP	2	-3.292	0.124	0.575	3
BADRDI	2	0.31	0.42	0.548	0
Pancreas/Bil tract					
NP	1	0.79	0.276	0.519	0
BADRDI	1	-0.154	0.416	0.515	0
NP	2	-1.733	0.237	0.499	0
BADRDI	2	0.6	0.287	0.494	0
Respir/Thorac					
OP	1	0.789	0.072	0.542	5
NP	1	1.152	0.155	0.534	4

CHAPITRE 5

BADRDI	1	-0.191	0.375	0.527	4
OP	2	-0.023	0.44	0.501	0
NP	2	-0.023	0.44	0.501	0
BADRDI	2	0.164	0.423	0.504	0
Bladder/Uroth					
NP	1	-0.383	0.431	0.501	0
BADRDI	1	-0.311	0.365	0.509	4
NP	2	-3.508	0.207	0.471	4
BADRDI	2	1.016	0.373	0.46	0

TABLE 5.7 – Prédiction des ajustements du modèle de Cox et détails gain/perte du pouvoir discriminant. Table du CHU de Boredaux.

Survival time difference and LogRank test significance details by quartile pair																					
Metric	Line	p-value	(-)	1 st vs 2 nd			1 st vs 3 rd			1 st vs 4 th			2 nd vs 3 rd			2 nd vs 4 th			3 rd vs 4 th		
				Method	p-value	(-)	Method														
Autoimm/Inflam																					
NP	1	0.3162	-73	AUC	0.0338	-140	AUC	0.0015	-288	AUC	0.1996	-67	AUC	0.0119	-215	AUC	0.1996	-148	AUC		
OP	1	0.0552	-14	AUC	0.0022	-215	AUC	0.0022	-270	AUC	0.0023	-256	AUC	0.952	-55	AUC	0.952	-55	AUC		
BADRDI	1	0.0606	128	AUC	0.0013	224	AUC	0.0606	159	AUC	0.1606	96	AUC	0.9352	30	AUC	0.1606	-66	AUC		
NP	2	0.4543	86	AUC	0.7181	3	AUC	0.7181	-55	AUC	0.2351	-83	AUC	0.2351	-141	AUC	0.8977	-58	AUC		
OP	2	0.7051	85	AUC	0.7202	75	AUC	0.8349	-31	AUC	0.7848	-10	AUC	0.7051	-116	AUC	0.7051	-106	AUC		
BADRDI	2	0.9877	-10	AUC	0.9463	32	AUC	0.5324	-83	AUC	0.9463	42	AUC	0.5324	-73	AUC	0.5324	-115	AUC		
Colon																					
NP	1	0.1181	266	Median	0.5834	-148	Median	0.1181	-300	Median	0.0695	-414	Median	0.004	-567	Median	0.2651	-152	Median		
OP	1	0.6368	83	Median	0.6269	40	Median	0.0083	-382	Median	0.8492	-42	Median	0.0032	-466	Median	0.0028	-423	Median		
BADRDI	1	0.1721	209	Median	0.1901	136	Median	0.8767	-24	Median	0.8422	-72	Median	0.1721	-234	Median	0.1802	-161	Median		
NP	2	0.2773	-212	Median	0.0029	-292	Median	0.9537	22	Median	0.0954	-80	Median	0.2733	234	Median	0.0029	314	Median		
OP	2	0.6949	-220	Median	0.1152	-275	Median	0.561	-129	Median	0.2435	-55	Median	0.6949	91	Median	0.3284	146	Median		
BADRDI	2	0.3958	28	Median	0.4622	-16	Median	0.1465	168	Median	0.1465	-44	Median	0.4622	140	Median	0.0569	184	Median		
Stomach																					
NP	1	0.6369	-210	Median	0.7229	-18	Median	0.6369	-206	Median	0.7229	192	Median	0.8941	4	Median	0.7229	-188	Median		
OP	1	0.9848	15	Median	0.9848	-8	Median	0.5318	-256	Median	0.9848	-24	Median	0.5318	-271	Median	0.57	-248	Median		
BADRDI	1	0.1965	-165	Median	0.1478	-116	Median	0.9626	91	Median	0.8522	49	Median	0.1965	256	Median	0.1478	207	Median		
NP	2	0.4729	-148	Median	0.4729	-59	Median	0.9114	-26	Median	0.9114	89	Median	0.4729	122	Median	0.4729	32	Median		
OP	2	0.3323	-160	Median	0.4248	-160	Median	0.9586	-8	Median	0.8345	0	Median	0.3323	152	Median	0.4248	153	Median		
BADRDI	2	0.9731	32	Median	0.5604	-2	Median	0.3516	211	Median	0.5604	-34	Median	0.3516	179	Median	0.1823	212	Median		
Lymph/Hemat																					
NP	1	0.1839	-168	AUC	9e-04	-922	Median	0.0123	-816	Median	0.0676	-922	Median	0.2426	-816	Median	0.4046	106	Median		
OP	1	0.0283	-210	AUC	5e-04	-740	Median	3e-04	-948	Median	0.2209	-740	Median	0.1839	-948	Median	0.8067	-208	Median		
BADRDI	1	0.0944	14	Median	0.8218	-558	Median	0.3655	-819	Median	0.0718	-572	Median	0.0102	-833	Median	0.4636	-261	Median		
NP	2	0.9072	-38	AUC	0.4845	-917	Median	0.6355	67	AUC	0.4845	-917	Median	0.6355	106	AUC	0.2993	917	Median		
OP	2	0.0754	-920	Median	0.2886	-605	Median	0.6344	57	AUC	0.4046	316	Median	0.0376	920	Median	0.1572	605	Median		
BADRDI	2	0.7909	163	Median	0.7909	-736	Median	0.7909	163	Median	0.7909	-899	Median	0.8069	84	AUC	0.7909	899	Median		

TABLE 5.9 – HEGP : Détails du pouvoir discriminant (1/3)

Survival time difference and LogRank test significance details by quartile pair																				
Metric	Line	p-value	1 st vs 2 nd			1 st vs 3 rd			1 st vs 4 th			2 nd vs 3 rd			2 nd vs 4 th			3 rd vs 4 th		
			(-)	Method	p-value	(-)	Method													
ENT																				
NP	1	0.76667	-167	Median	0.7035	-389	Median	0.7035	-310	Median	0.7035	-222	Median	0.76667	-144	Median	0.76667	78	Median	
OP	1	0.7002	53	Median	0.7002	-266	Median	0.7636	-196	Median	0.7002	-318	Median	0.7002	-248	Median	0.7002	70	Median	
BADRDI	1	0.54333	213	Median	0.0836	-236	Median	0.0013	-506	Median	0.02	-448	Median	1e-04	-718	Median	0.0836	-270	Median	
NP	2	0.4042	19	Median	0.9545	-168	Median	2e-04	-320	Median	0.4042	-188	Median	0	-340	Median	4e-04	-152	Median	
OP	2	0.1043	-138	Median	0.8464	30	Median	0	-344	Median	0.1603	168	Median	0.0015	-206	Median	0	-374	Median	
BADRDI	2	0.4989	161	Median	0.2855	-178	Median	0.0295	-146	Median	0.0977	-339	Median	0.005	-307	Median	0.2855	32	Median	
Ovary																				
NP	1	0.3983	-435	Median	0.019	-878	Median	0.0035	-935	Median	0.1162	-443	Median	0.0239	-500	Median	0.4854	-57	Median	
OP	1	0.9503	-16	AUC	0.002	-969	Median	0.002	-935	Median	0.002	-969	Median	0.002	-935	Median	0.9503	34	Median	
BADRDI	1	0.6046	286	Median	0.4446	482	Median	0.6909	66	Median	0.6046	196	Median	0.5058	-220	Median	0.3905	-416	Median	
NP	2	0.1117	-665	Median	7e-04	-1176	Median	0.0022	-1150	Median	0.0581	-510	Median	0.1117	-484	Median	0.7133	26	Median	
OP	2	0.0309	-796	Median	0	-1160	Median	1e-04	-1215	Median	0.0325	-364	Median	0.0534	-419	Median	0.8839	-56	Median	
BADRDI	2	0.5967	416	Median	0.9214	130	Median	0.0585	-360	Median	0.5967	-285	Median	0.0149	-776	Median	0.0475	-490	Median	
Pancreas/Bil tract																				
NP	1	0.00112	-202	Median	0.019	-194	Median	0.014	-151	Median	0.3891	9	Median	0.3891	52	Median	0.9128	42	Median	
OP	1	0.0577	-193	Median	0.1645	-196	Median	0.0577	-209	Median	0.6658	-2	Median	0.8511	-16	Median	0.6546	-14	Median	
BADRDI	1	0.8533	98	Median	0.8533	14	Median	0.8533	86	Median	0.8533	-84	Median	0.8533	-12	Median	0.8533	73	Median	
NP	2	0.6828	-34	Median	0.6828	-53	Median	0.6828	-20	Median	0.6828	-19	Median	0.6828	14	Median	0.6828	34	Median	
OP	2	0.0232	114	Median	0.0101	206	Median	0.0028	272	Median	0.9024	92	Median	0.6549	158	Median	0.6549	66	Median	
BADRDI	2	0.5576	5	Median	0.4281	152	Median	0.0963	21	Median	0.1689	147	Median	0.287	16	Median	0.0072	-131	Median	
Respir/Thorac																				
NP	1	0.8934	-69	Median	0.0491	-162	Median	0	-227	Median	0.06	-93	Median	0	-158	Median	0.0334	-65	Median	
OP	1	0.0011	121	Median	0.0421	-147	Median	0	-209	Median	0	-268	Median	0	-330	Median	0.034	-62	Median	
BADRDI	1	0.8881	-21	Median	0.8881	4	Median	0.209	-74	Median	0.8881	25	Median	0.209	-53	Median	0.209	-78	Median	
NP	2	0.2497	-4	Median	0.862	-56	Median	0.2497	-104	Median	0.2497	-52	Median	0.0219	-100	Median	0.2497	-47	Median	
OP	2	0.0615	114	Median	0.5706	-10	Median	0.0615	-122	Median	0.2113	-124	Median	2e-04	-236	Median	0.0227	-112	Median	
BADRDI	2	0.0191	180	Median	0.1038	70	Median	0.7253	27	Median	0.3886	-110	Median	0.0191	-153	Median	0.144	-42	Median	

TABLE 5.11 – HEGP : Détails du pouvoir discriminant (2/3)

Survival time difference and LogRank test significance details by quartile pair																
Metric	Line	1 st vs 2 nd			1 st vs 3 rd			1 st vs 4 th			2 nd vs 3 rd			2 nd vs 4 th		
		p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method
Breast																
NP	1	1e-04	120	AUC	0.0222	-177	AUC	0	-1054	Median	0	-296	AUC	0	-1054	Median
OP	1	0.032	58	AUC	0.0043	-181	AUC	0	-1047	Median	0	-239	AUC	0	-1047	Median
BADRDI	1	0.0086	166	AUC	0.3596	0	AUC	0	-1102	Median	3e-04	-166	AUC	0	-1102	Median
NP	2	0.2222	73	AUC	0.0663	-364	Median	0	-1394	Median	0.0023	-364	Median	0	-1394	Median
OP	2	0.0705	134	AUC	0.0322	-689	Median	0	-1378	Median	1e-04	-689	Median	5e-04	-689	Median
BADRDI	2	0.0035	177	Median	0.6057	46	Median	0.0017	-1124	Median	0.0125	-131	Median	4e-04	-1170	Median
Bladder/Uroth																
NP	1	0.1119	-300	Median	0.8734	162	Median	0.0101	-523	Median	0.1325	462	Median	0.2259	-223	Median
OP	1	0.0571	843	Median	0.0452	-353	Median	4e-04	-570	Median	2e-04	-1196	Median	0	-1413	Median
BADRDI	1	0.707	-64	Median	0.0061	-492	Median	0.707	138	Median	0.0077	-428	Median	0.9286	202	Median
NP	2	0.0363	-370	Median	0.0363	-344	Median	1e-04	-576	Median	0.0363	-206	Median	0.0363	-232	Median
OP	2	0.9468	-34	Median	0.0026	-418	Median	0.001	-492	Median	0.0022	-383	Median	9e-04	-457	Median
BADRDI	2	0.8227	-124	Median	0.0432	-142	Median	0.4786	-218	Median	0.0474	-19	Median	0.4786	-94	Median
Vulva																
NP	1	0.1066	424	Median	0.5265	-228	Median	0.253	-276	Median	0.0387	-652	Median	0.0071	-700	Median
OP	1	0.4154	138	Median	0.9154	-260	Median	0.4154	-283	Median	0.4154	-398	Median	0.1453	-421	Median
BADRDI	1	0.8281	229	Median	0.8281	195	Median	0.0647	-209	Median	0.8281	-34	Median	0.0176	-438	Median
NP	2	0.7153	231	Median	0.7153	-94	Median	0.9736	202	Median	0.7153	-325	Median	0.7153	-28	Median
OP	2	0.9704	40	Median	0.9704	-236	Median	0.9704	108	Median	0.9704	-276	Median	0.9704	68	Median
BADRDI	2	0.1581	369	Median	0.3772	308	Median	0.3772	262	Median	0.3714	-61	Median	0.3714	-107	Median

TABLE 5.13 – HEGP : Détails du pouvoir discriminant (3/3)

Survival time difference and LogRank test significance details by quartile pair																				
Metric	Line	p-value	1 st vs 2 nd			1 st vs 3 rd			1 st vs 4 th			2 nd vs 3 rd			2 nd vs 4 th			3 rd vs 4 th		
			(-)	Method	p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method	
Autoimm/Inflam																				
NP	1	0.0705	-61	AUC	0.3013	-41	AUC	0.0705	-41	AUC	0.5022	19	AUC	0.9435	20	AUC	0.5022	1	AUC	
BADRDI	1	0.5208	-11	AUC	0.5208	-24	AUC	0.5208	4	AUC	0.5208	-13	AUC	0.9246	15	AUC	0.5208	28	AUC	
NP	2	0.4692	43	AUC	0.7643	12	AUC	0.4692	38	AUC	0.5171	-30	AUC	0.8694	-4	AUC	0.5171	26	AUC	
BADRDI	2	0.9762	15	AUC	0.9762	16	AUC	0.9762	28	AUC	0.9762	1	AUC	0.9762	14	AUC	0.9762	12	AUC	
Colon																				
NP	1	0.0335	102	Median	6e-04	789	Median	6e-04	782	Median	0.1104	687	Median	0.0835	680	Median	0.5907	-7	Median	
BADRDI	1	0.6329	143	Median	0.5416	498	Median	0.5416	247	Median	0.5416	355	Median	0.6329	104	Median	0.6329	-251	Median	
NP	2	0.0046	138	Median	1e-04	300	Median	0	898	Median	0.2621	162	Median	0.0398	760	Median	0.4107	598	Median	
BADRDI	2	0.646	104	Median	0.0351	259	Median	0.0122	669	Median	0.0325	155	Median	0.0122	565	Median	0.646	410	Median	
Melanoma																				
NP	1	0.0163	101	AUC	0	-1450	Median	0	-1450	Median	0	-1450	Median	0	-1450	Median	0.1917	1	Median	
BADRDI	1	0.9	-606	Median	0.3589	166	Median	0.3589	80	Median	0.3589	772	Median	0.3589	686	Median	0.9238	-86	Median	
NP	2	0.8242	59	AUC	0.8242	39	AUC	0	-1439	Median	0.9769	-21	AUC	0	-1439	Median	0	-1439	Median	
BADRDI	2	0.5836	-305	Median	0.2628	141	AUC	0.5836	62	AUC	0.1516	305	Median	0.3866	305	Median	0.3866	-78	AUC	
Liver																				
NP	1	0.1665	608	Median	0.1666	601	Median	0.0073	901	Median	0.87	-8	Median	0.1665	292	Median	0.1665	300	Median	
BADRDI	1	0.6244	-40	Median	0.2024	-366	Median	0.0453	546	Median	0.403	-325	Median	0.0244	587	Median	0.0012	912	Median	
NP	2	0.6015	200	Median	0.6015	851	Median	0.5722	1086	Median	0.9927	651	Median	0.5722	886	Median	0.5722	236	Median	
BADRDI	2	0.9895	-80	Median	0.0941	1020	Median	0.6604	708	Median	0.0941	1100	Median	0.6604	788	Median	0.3641	-312	Median	
Myeloma																				
NP	1	0.4701	-2	AUC	0.4701	-17	AUC	0	258	AUC	0.9232	-15	AUC	2e-04	260	AUC	2e-04	274	AUC	
BADRDI	1	0.0085	248	AUC	0.0896	110	AUC	0	356	AUC	0.3834	-138	AUC	0.0084	108	AUC	0.0012	246	AUC	
NP	2	0.2373	146	AUC	0.1207	194	AUC	0.1207	180	AUC	0.7649	49	AUC	0.7649	34	AUC	0.95	-14	AUC	
BADRDI	2	0.0979	173	AUC	0.1598	151	AUC	1e-04	356	AUC	0.7095	-22	AUC	0.0232	183	AUC	0.011	205	AUC	
Lymph/Hemat																				
NP	1	1e-04	-103	AUC	0	-223	AUC	0	-252	AUC	1e-04	-121	AUC	0	-150	AUC	0.2042	-29	AUC	
BADRDI	1	0.366	-15	AUC	0	-204	AUC	0	-170	AUC	0	-189	AUC	0	-155	AUC	0.2685	34	AUC	
NP	2	0.0843	55	AUC	0.3199	-27	AUC	6e-04	-144	AUC	0.0092	-82	AUC	0	-199	AUC	0.0092	-117	AUC	
BADRDI	2	0.7466	-1	AUC	0.5795	-20	AUC	0.0026	-133	AUC	0.6827	-19	AUC	0.0035	-131	AUC	0.0136	-112	AUC	

TABLE 5.15 – Bordeaux : Détails du pouvoirs discriminant (1/3)

Survival time difference and LogRank test significance details by quartile pair																					
Metric	Line	p-value	(-)	1 st vs 2 nd			1 st vs 3 rd			1 st vs 4 th			2 nd vs 3 rd			2 nd vs 4 th			3 rd vs 4 th		
				p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method			
Neurology																					
NP	1	0.4404	-7	AUC	0.4404	59	AUC	0.2207	84	AUC	0.2207	66	AUC	0.0683	91	AUC	0.4404	25	AUC		
BADRDI	1	0.2148	38	AUC	0.017	104	AUC	0.1946	87	AUC	0.2148	66	AUC	0.7985	49	AUC	0.2612	-17	AUC		
NP	2	0.3458	49	AUC	0.4322	28	AUC	0.0108	126	AUC	0.7722	-20	AUC	0.0666	78	AUC	0.0535	98	AUC		
BADRDI	2	0.7853	-39	AUC	0.2822	43	AUC	0.0532	86	AUC	0.2417	82	AUC	0.0532	126	AUC	0.2961	44	AUC		
ENT																					
NP	1	0	1443	Median	0	1443	Median	0	1443	Median	0	334	AUC	0.0005	196	AUC	0.2009	-138	AUC		
BADRDI	1	0.8001	-23	AUC	0.1125	-254	Median	0.1882	-176	AUC	0.1125	-254	Median	0.1954	-153	AUC	0.8001	254	Median		
NP	2	0.2274	138	Median	0.0347	980	Median	0.001	1484	Median	0.2274	842	Median	0.0144	1346	Median	0.1934	504	Median		
BADRDI	2	0.3918	1013	Median	0.5009	-26	Median	0.551	299	Median	0.0876	-1039	Median	0.551	-714	Median	0.2908	325	Median		
Pancreas/Bil tract																					
NP	1	0.9656	15	Median	0.3544	36	Median	0.3544	-5	Median	0.3544	21	Median	0.3544	-20	Median	0.1046	-41	Median		
BADRDI	1	0.0662	148	Median	0.0662	115	Median	0.9505	95	Median	0.9697	-34	Median	0.0662	-54	Median	0.0662	-20	Median		
NP	2	0.1838	160	Median	0.3803	106	Median	0.0705	158	Median	0.4522	-54	Median	0.3803	-2	Median	0.2045	52	Median		
BADRDI	2	0.5642	138	Median	0.4382	132	Median	0.4382	69	Median	0.7664	-6	Median	0.7664	-69	Median	0.7664	-63	Median		
Respir/Thorac																					
NP	1	0.0848	-472	Median	1e-04	-670	Median	0.4012	846	Median	0.0242	-198	Median	0.0242	1318	Median	0	1516	Median		
OP	1	0.0144	-697	Median	0	-726	Median	0.3368	792	Median	0.0063	-28	Median	0.0063	1488	Median	0	1517	Median		
BADRDI	1	0.0218	667	Median	0.982	200	Median	0.0316	1100	Median	0.0218	-468	Median	0.0862	432	Median	0.0218	900	Median		
NP	2	0.8153	-199	Median	0.8554	138	Median	0.9394	60	Median	0.8554	338	Median	0.8153	258	Median	0.8554	-79	Median		
OP	2	0.8153	-199	Median	0.8554	138	Median	0.9394	60	Median	0.8554	338	Median	0.8153	258	Median	0.8554	-79	Median		
BADRDI	2	0.4824	66	Median	0.8894	128	Median	0.2451	406	Median	0.4824	62	Median	0.4824	340	Median	0.2451	278	Median		
Bladder/Uroth																					
NP	1	0.6336	-96	AUC	0.6336	-50	AUC	0.8848	21	AUC	0.9268	46	AUC	0.6336	117	AUC	0.6336	71	AUC		
BADRDI	1	0.5754	26	AUC	5e-04	-335	AUC	7e-04	-296	AUC	1e-04	-361	AUC	1e-04	-322	AUC	0.8399	40	AUC		
NP	2	0.0443	461	AUC	0.6918	45	AUC	0.0318	548	AUC	0.0482	-416	AUC	0.6918	87	AUC	0.0318	503	AUC		
BADRDI	2	0.5006	-159	AUC	0.5006	-249	AUC	0.9255	66	AUC	0.9244	-90	AUC	0.5006	225	AUC	0.5006	315	AUC		

TABLE 5.17 – Bordeaux : Détails du pouvoir discriminant (2/3)

Survival time difference and LogRank test significance details by quartile pair																				
Metric	Line	p-value	1 st vs 2 nd			1 st vs 3 rd			1 st vs 4 th			2 nd vs 3 rd			2 nd vs 4 th			3 rd vs 4 th		
			(-)	Method	p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method	p-value	(-)	Method	
Neurology																				
NP	1	0.585	4	AUC	0.1802	71	AUC	0.0669	100	AUC	0.0669	66	AUC	0.038	95	AUC	0.585	29	AUC	
OP	1	0.585	4	AUC	0.1802	71	AUC	0.0669	100	AUC	0.0669	66	AUC	0.038	95	AUC	0.585	29	AUC	
BADRDI	1	0.7974	-7	AUC	0.0352	85	AUC	0.1341	77	AUC	0.0379	93	AUC	0.1794	85	AUC	0.3812	-8	AUC	
NP	2	0.2433	53	AUC	0.3177	33	AUC	0.0057	129	AUC	0.7729	-20	AUC	0.0644	76	AUC	0.0516	95	AUC	
OP	2	0.2433	53	AUC	0.3177	33	AUC	0.0057	129	AUC	0.7729	-20	AUC	0.0644	76	AUC	0.0516	95	AUC	
BADRDI	2	0.8528	-16	AUC	0.3308	30	AUC	0.0507	90	AUC	0.3707	46	AUC	0.0507	106	AUC	0.2875	60	AUC	
ENT																				
NP	1	1e-04	1310	Median	0	1386	Median	0	1386	Median	0	76	Median	0.004	76	Median	0.1824	-128	AUC	
OP	1	1e-04	1310	Median	0	1386	Median	0	1386	Median	0	76	Median	0.004	76	Median	0.1824	-128	AUC	
BADRDI	1	0.4217	-100	AUC	0.0294	-606	Median	0.1532	-166	AUC	0.175	-606	Median	0.4217	-66	AUC	0.4217	606	Median	
NP	2	0.9769	93	Median	0.3455	853	Median	0.1952	1244	Median	0.3455	760	Median	0.1952	1151	Median	0.594	391	Median	
OP	2	0.9769	93	Median	0.3455	853	Median	0.1952	1244	Median	0.3455	760	Median	0.1952	1151	Median	0.594	391	Median	
BADRDI	2	0.4371	1315	Median	0.6092	38	Median	0.6092	270	Median	0.2053	-1277	Median	0.5824	-1045	Median	0.4371	232	Median	
Pancreas/Bil tract																				
NP	1	0.6407	-28	Median	0.1878	46	Median	0.1878	-36	Median	0.1259	75	Median	0.3583	-8	Median	0.0128	-83	Median	
OP	1	0.6407	-28	Median	0.1878	46	Median	0.1878	-36	Median	0.1259	75	Median	0.3583	-8	Median	0.0128	-83	Median	
BADRDI	1	0.1426	138	Median	0.1426	116	Median	0.6052	90	Median	0.8426	-22	Median	0.2544	-48	Median	0.2278	-26	Median	
NP	2	0.9302	56	Median	0.9302	95	Median	0.2021	121	Median	0.9302	40	Median	0.2021	66	Median	0.2021	26	Median	
OP	2	0.9302	56	Median	0.9302	95	Median	0.2021	121	Median	0.9302	40	Median	0.2021	66	Median	0.2021	26	Median	
BADRDI	2	0.8622	108	Median	0.7916	104	Median	0.7916	54	Median	0.8202	-4	Median	0.7916	-54	Median	0.7916	-50	Median	
Polyarth																				
NP	1	0.2517	-111	AUC	0.2517	-105	AUC	0.8611	-13	AUC	0.8992	6	AUC	0.2517	98	AUC	0.2517	92	AUC	
OP	1	0.2517	-111	AUC	0.2517	-105	AUC	0.8611	-13	AUC	0.8992	6	AUC	0.2517	98	AUC	0.2517	92	AUC	
BADRDI	1	0.4365	-66	AUC	0.4365	-35	AUC	0.6365	-14	AUC	0.9635	31	AUC	0.6365	52	AUC	0.6365	21	AUC	
NP	2	0.1708	-103	AUC	0.4867	-46	AUC	0.3968	-53	AUC	0.3968	57	AUC	0.3968	50	AUC	0.7427	-7	AUC	
OP	2	0.1708	-103	AUC	0.4867	-46	AUC	0.3968	-53	AUC	0.3968	57	AUC	0.3968	50	AUC	0.7427	-7	AUC	
BADRDI	2	0.6876	-65	AUC	0.6998	-49	AUC	0.7092	-9	AUC	0.6998	17	AUC	0.6876	56	AUC	0.7092	40	AUC	

TABLE 5.19 – Bordeaux : Détails du pouvoir discriminant (3/3)

BIBLIOGRAPHIE

- [1] Robert D ABBOTT. « Logistic regression in survival analysis ». In : *American journal of epidemiology* 121.3 (1985), p. 465-471.
- [2] Soumeya L ACHOUR et al. « A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development ». In : *Journal of the American Medical Informatics Association* 8.4 (2001), p. 351-360.
- [3] Najia AHMADI et al. « OMOP CDM can facilitate data-driven studies for cancer prediction : a systematic review ». In : *International Journal of Molecular Sciences* 23.19 (2022), p. 11834.
- [4] James F ALLEN. « Maintaining knowledge about temporal intervals ». In : *Communications of the ACM* 26.11 (1983), p. 832-843.
- [5] AMERICAN CANCER SOCIETY. *How Chemotherapy Drugs Work*. Accessed : May 3, 2023. 2023. URL : <https://www.cancer.org/cancer/managing-cancer/treatment-types/chemotherapy/how-chemotherapy-drugs-work.html> (visité le 03/05/2023).
- [6] Larissa Costa AMORIM et Renata D'Alpino PEIXOTO. « Should we still be using bolus 5-FU prior to infusional regimens in gastrointestinal cancers ? A practical review ». In : *International Cancer Conference Journal*. T. 11. 1. Springer. 2022, p. 2-5.
- [7] Sylvain ARLOT et Alain CELISSE. « A survey of cross-validation procedures for model selection ». In : *Statistics Surveys* (2010).
- [8] Wayne J ASTON et al. « A systematic investigation of the maximum tolerated dose of cytotoxic chemotherapy with and without supportive care in mice ». In : *BMC cancer* 17 (2017), p. 1-10.
- [9] Benu ATRI et Olivier LICHTARGE. « Sequence Alignment ». In : *Bioinformatics : Sequences, Structures, Phylogeny* (2018), p. 47-69.
- [10] Paul AVILLACH et al. « Design and validation of an automated method to detect known adverse drug reactions in MEDLINE : a contribution from the EU-ADR project ». In : *Journal of the American Medical Informatics Association* 20.3 (2013), p. 446-452.
- [11] Franz BAADER. *The description logic handbook : Theory, implementation and applications*. Cambridge university press, 2003.
- [12] Vinod P BALACHANDRAN et al. « Nomograms in oncology : more than meets the eye ». In : *The lancet oncology* 16.4 (2015), e173-e180.
- [13] Lodovico BALDUCCI et Martine EXTERMANN. « Cancer chemotherapy in the older patient : what the medical oncologist needs to know ». In : *Cancer : Interdisciplinary International Journal of the American Cancer Society* 80.7 (1997), p. 1317-1322.

- [14] Nesrine BANNOUR et al. « Event-independent temporal positioning : application to French clinical text ». In : *22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Association for Computational Linguistics. 2023, p. 191-205.
- [15] Sotiris BATSAKIS et al. « Temporal representation and reasoning in OWL 2 ». In : *Semantic Web 8.6* (2017), p. 981-1000.
- [16] Rimma BELENKAYA et al. « Extending the OMOP common data model and standardized vocabularies to support observational cancer research ». In : *JCO Clinical Cancer Informatics* 5 (2021).
- [17] Sébastien BENZEKRY et al. « Metronomic reloaded : Theoretical models bringing chemotherapy into the era of precision medicine ». In : *Seminars in Cancer Biology*. T. 35. Elsevier. 2015, p. 53-61.
- [18] James BERGSTRA et al. « Algorithms for hyper-parameter optimization ». In : *Advances in neural information processing systems* 24 (2011).
- [19] Tim BERNERS-LEE, James HENDLER et Ora LASSILA. « The Semantic Web ». In : *Scientific American* 284.5 (2001). JSTOR 26059207, p. 34-43. URL : <https://www.jstor.org/stable/26059207>.
- [20] José Luiz B BEVILACQUA et al. « Doctor, what are my chances of having a positive sentinel node ? A validated nomogram for risk estimation ». In : *Journal of clinical oncology* 25.24 (2007), p. 3670-3679.
- [21] Paul BEYNON-DAVIES. « Database Systems and Electronic Business ». In : *Database Systems*. Springer, 2004, p. 62-75.
- [22] Edith BOYD. « The growth of the surface area o ! the human body. » In : (1935).
- [23] Melissa BRACKMANN et al. « Comparison of first-line chemotherapy regimens for ovarian carcinosarcoma : a single institution case series and review of the literature ». In : *BMC cancer* 18 (2018), p. 1-7.
- [24] Daniel BREADNER et al. « The influence of adjuvant chemotherapy dose intensity on overall survival in resected colon cancer : a multicentered retrospective analysis ». In : *BMC cancer* 22.1 (2022), p. 1119.
- [25] Norman BRESLOW. « A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship ». In : *Biometrika* 57.3 (1970), p. 579-594.
- [26] Norman BRESLOW. « Covariance analysis of censored survival data ». In : *Biometrics* (1974), p. 89-99.
- [27] Lauren C BYLSMA et al. « Chemotherapy Relative Dose Intensity, Overall Survival, and Hematologic Toxicity in Solid-Tumor Cancer Patients : A Literature Review and Meta-Analysis ». In : *Blood* 136 (2020), p. 32-33.

- [28] Institut National du CANCER. *Activité hospitalière liée aux traitements du cancer en France*. 2021. URL : <https://www.e-cancer.fr/Professionnels-de-sante/Les-chiffres-du-cancer-en-France/Activite-hospitaliere> (visité le 03/05/2023).
- [29] Kevin J CARROLL. « On the use and utility of the Weibull model in the analysis of survival data ». In : *Controlled clinical trials* 24.6 (2003), p. 682-701.
- [30] Stefano CASCINU, Elena DEL FERRO et Giuseppina CATALANO. « Toxicity and therapeutic response to chemotherapy in patients aged 70 years or older with advanced cancer ». In : *American journal of clinical oncology* 19.4 (1996), p. 371-374.
- [31] CEPD. *Toxicité des traitements : Critères de toxicité*. http://www.cepd.fr/CUSTOM/CEPD_toxicite.pdf. Accessed : 2024-06-10. 2024.
- [32] Ritesh CHANDRA et al. « Semantic web-based diagnosis and treatment of vector-borne diseases using SWRL rules ». In : *Knowledge-Based Systems* 274 (2023), p. 110645.
- [33] Etienne CHATELUT et al. « Prediction of carboplatin clearance from standard morphological and biological patient characteristics ». In : *JNCI : Journal of the National Cancer Institute* 87.8 (1995), p. 573-580.
- [34] CHEBI. *ChEBI : Chemical Entities of Biological Interest*. <https://www.ebi.ac.uk/chebi/>. Accessed : 2024-06-10. 2024.
- [35] Henry W CHEN et al. « Representation of time-relevant common data elements in the cancer data standards repository : statistical evaluation of an ontological approach ». In : *JMIR medical informatics* 6.1 (2018), e8175.
- [36] Yifei CHEN et al. « A gradient boosting algorithm for survival analysis via direct optimization of concordance index ». In : *Computational and mathematical methods in medicine* 2013.1 (2013), p. 873595.
- [37] Edward CHU et Vincent T DeVITA Jr. *Physicians' Cancer Chemotherapy Drug Manual* 2024. Jones & Bartlett Learning, 2023.
- [38] Taane G CLARK et al. « Survival analysis part I : basic concepts and first analyses ». In : *British journal of cancer* 89.2 (2003), p. 232-238.
- [39] Ronan COLLOBERT et Jason WESTON. « A unified architecture for natural language processing : Deep neural networks with multitask learning ». In : *Proceedings of the 25th international conference on Machine learning*. 2008, p. 160-167.
- [40] COLLÈGE NATIONAL DE PHARMACOLOGIE MÉDICALE. *PharmaMedicale*. 2023. URL : <https://pharmacomedicale.org/medicaments/par-specialites/item/anticancereux-les-points-essentiels> (visité le 03/05/2023).

- [41] Sébastien COSSIN et al. « Linkage of Hospital Records and Death Certificates by a Search Engine and Machine Learning ». In : *JAMIA Open* 4.1 (mars 2021), ooab005. ISSN : 2574-2531. doi : 10.1093/jamiaopen/ooab005. eprint : <https://academic.oup.com/jamiaopen/article-pdf/4/1/ooab005/36416919/ooab005.pdf>. URL : <https://doi.org/10.1093/jamiaopen/ooab005>.
- [42] Sébastien COSSIN et al. « Romedi : an open data source about French drugs on the semantic web ». In : *MEDINFO 2019 : Health and Wellbeing e-Networks for All*. IOS Press, 2019, p. 79-82.
- [43] Sébastien COSSIN et al. « IAM at CLEF eHealth 2018 : Concept Annotation and Coding in French Death Certificates ». In : *arXiv :1807.03674 [cs]* (juill. 2018). arXiv : 1807.03674. URL : <http://arxiv.org/abs/1807.03674> (visité le 11/07/2018).
- [44] David R Cox. « Regression models and life-tables ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 34.2 (1972), p. 187-202.
- [45] Simon Cox et Chris LITTLE. *Time Ontology in OWL*. World Wide Web Consortium (W3C) Candidate Recommendation Draft. OGC Document Number : OGC 16-071r3. 2022. URL : <https://www.w3.org/TR/2022/CRD-owl-time-20221115/>.
- [46] Amit DANG. « Real-world evidence : a primer ». In : *Pharmaceutical medicine* 37.1 (2023), p. 25-36.
- [47] Amanda DELGADO et Achuta Kumar GUDDATI. « Clinical endpoints in oncology-a primer ». In : *American journal of cancer research* 11.4 (2021), p. 1121.
- [48] Maxime DELMAS. « Construire, exploiter et étendre un graphe de connaissances pour l'étude des liens entre métabolisme et santé ». Thèse de doct. Université Paul Sabatier-Toulouse III, 2022.
- [49] Maxime DELMAS et al. « building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases ». In : *Bioinformatics* 37.21 (2021), p. 3896-3904.
- [50] Neelima DENDULURI et al. « Dose delays, dose reductions, and relative dose intensity in patients with cancer who received adjuvant or neoadjuvant chemotherapy in community oncology practices ». In : *Journal of the National Comprehensive Cancer Network* 13.11 (2015), p. 1383-1393.
- [51] Karl-Matthias DEPPERMAN. « Influence of age and comorbidities on the chemo-therapeutic management of lung cancer ». In : *Lung Cancer* 33 (2001), S115-S120.
- [52] William DIGAN et al. « PyMedExt, un couteau suisse pour le traitement des textes médicaux ». In : *Paris : AFIA-TLH/ATALA* (2021).

- [53] Martin DRANCE et al. « Pre-Trained Embeddings for Enhancing Multi-Hop Reasoning ». In : *International Joint Conference on Artificial Intelligence 2023 Workshop on Knowledge-Based Compositional Generalization*. 2023.
- [54] DRON. *Drug Ontology (DRON)*. <https://bioportal.bioontology.org/ontologies/DRON>. Accessed : 2024-06-10. 2024.
- [55] DRUGBANK. *DrugBank : The Drug Database*. <https://go.drugbank.com/>. Accessed : 2024-06-10. 2024.
- [56] Thomas EFFERTH et Manfred VOLM. « Pharmacogenetics for individualized cancer chemotherapy ». In : *Pharmacology & therapeutics* 107.2 (2005), p. 155-176.
- [57] Bradley EFRON. « The efficiency of Cox's likelihood function for censored data ». In : *Journal of the American statistical Association* 72.359 (1977), p. 557-565.
- [58] Bradley EFRON et Robert J TIBSHIRANI. *An introduction to the bootstrap*. Chapman et Hall/CRC, 1994.
- [59] Tome EFTIMOV, Barbara KOROUŠIĆ SELJAK et Peter KOROŠEC. « A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations ». In : *PLoS one* 12.6 (2017), e0179488.
- [60] Lisa EHRLINGER et Wolfram Wöss. « Towards a definition of knowledge graphs. » In : *SEMANTiCS (Posters, Demos, SuCESS)* 48.1-4 (2016), p. 2.
- [61] Computer ENGINEERING. *Chimio®*. Accessed : 2024-07-05. 2024. URL : <https://www.computer-engineering.fr/applications/chimio/>.
- [62] EUROPEAN SOCIETY FOR MEDICAL ONCOLOGY (ESMO). *ESMO Clinical Practice Guidelines*. Accessed : May 3, 2023. 2023. URL : <https://www.esmo.org/guidelines> (visité le 03/05/2023).
- [63] Usama FAYYAD, Gregory PIATETSKY-SHAPIRO et Padhraic SMYTH. « From data mining to knowledge discovery in databases ». In : *AI magazine* 17.3 (1996), p. 37-37.
- [64] Dieter FENSEL. *Spinning the Semantic Web : bringing the World Wide Web to its full potential*. MIT press, 2005.
- [65] Dieter FENSEL et al. *Knowledge graphs*. Springer, 2020.
- [66] Elizabeth FORD et al. « Extracting information from the text of electronic medical records to improve case detection : a systematic review ». In : *Journal of the American Medical Informatics Association* 23.5 (2016), p. 1007-1015.
- [67] J FRAISSE et al. « Optimal biological dose : a systematic review in cancer phase I clinical trials ». In : *BMC cancer* 21 (2021), p. 1-10.
- [68] Alexandre GALOPIN et al. « An ontology-based clinical decision support system for the management of patients with multiple chronic disorders ». In : *MEDINFO 2015 : eHealth-enabled Health*. IOS Press, 2015, p. 275-279.
- [69] Valentina GAMBARDELLA et al. « Personalized medicine : recent progress in cancer therapy ». In : *Cancers* 12.4 (2020), p. 1009.

- [70] Nicolas GARCELON. « Problématique des entrepôts de données textuelles : dr Warehouse et la recherche translationnelle sur les maladies rares ». Thèse de doct. Université Sorbonne Paris Cité, 2017.
- [71] Nicolas GARCELON et al. « Improving a full-text search engine : the importance of negation detection and family history context to identify cases in a biomedical data warehouse ». In : *Journal of the American Medical Informatics Association* 24.3 (2017), p. 607-613.
- [72] Olga GOLUBNITSCHAJA, Judita KINKROVA et Vincenzo COSTIGLIOLA. « Predictive, preventive and personalised medicine as the hardcore of ‘Horizon 2020’ : EPMA position paper ». In : *EPMA Journal* 5 (2014), p. 1-29.
- [73] Asunción GÓMEZ-PÉREZ, Mariano FERNÁNDEZ-LÓPEZ et Oscar CORCHO. *Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media, 2006.
- [74] Michelle GOODMAN. « Managing the side effects of chemotherapy. » In : *Seminars in oncology nursing*. T. 5. 2 Suppl 1. 1989, p. 29-52.
- [75] Dev GOYAL, Zeeshan SYED et Jenna WIENS. « Clinically meaningful comparisons over time : An approach to measuring patient similarity based on subsequence alignment ». In : *arXiv preprint arXiv :1803.00744* (2018).
- [76] Olivier GRAESSLIN et al. « Nomogram to predict subsequent brain metastasis in patients with metastatic breast cancer ». In : *Journal of clinical oncology* 28.12 (2010), p. 2032-2037.
- [77] Aimery de GRAMONT et al. « Randomized trial comparing monthly low-dose leucovorin and fluorouracil bolus with bimonthly high-dose leucovorin and fluorouracil bolus plus continuous infusion for advanced colorectal cancer : a French intergroup study. » In : *Journal of Clinical Oncology* 15.2 (1997), p. 808-815.
- [78] Richard GRAY et al. « Increasing the dose intensity of chemotherapy by more frequent administration or sequential scheduling : a patient-level meta-analysis of 37 298 women with early breast cancer in 26 randomised trials ». In : *The lancet* 393.10179 (2019), p. 1440-1452.
- [79] Cesare GRIDELLI et Frances A SHEPHERD. « Chemotherapy for elderly patients with non-small cell lung cancer : a review of the evidence ». In : *Chest* 128.2 (2005), p. 947-957.
- [80] Cesare GRIDELLI et al. « Chemotherapy for elderly patients with advanced non-small-cell lung cancer : the Multicenter Italian Lung Cancer in the Elderly Study (MILES) phase III randomized trial ». In : *Journal of the National Cancer Institute* 95.5 (2003), p. 362-372.

- [81] Nicola GUARINO, Daniel OBERLE et Steffen STAAB. « What Is an Ontology? » In : *Handbook on Ontologies*. Sous la dir. de Steffen STAAB et Rudi STUDER. Berlin, Heidelberg : Springer Berlin Heidelberg, 2009, p. 1-17. ISBN : 978-3-540-92673-3. DOI : [10.1007/978-3-540-92673-3_0](https://doi.org/10.1007/978-3-540-92673-3_0). URL : https://doi.org/10.1007/978-3-540-92673-3_0.
- [82] Peter S HALL et al. « Efficacy of reduced-intensity chemotherapy with oxaliplatin and capecitabine on quality of life and cancer control among older and frail patients with advanced gastroesophageal cancer : the GO2 phase 3 randomized clinical trial ». In : *JAMA oncology* 7.6 (2021), p. 869-877.
- [83] Frank E HARRELL et al. *Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis*. T. 608. Springer, 2001.
- [84] Mohsen HASSAN et al. « Extracting disease-symptom relationships by learning syntactic patterns from dependency graphs ». In : *BioNLP 15*. 2015, p. 184.
- [85] HAUTE AUTORITÉ DE SANTÉ (HAS). *Résultat de recherche : Recommandations et référentiels*. Accessed : May 3, 2023. 2023. URL : https://www.has-sante.fr/jcms/fc_2875171/fr/resultat-de-recherche?FACET_THEME=c_64705%252Fc_1151857&types=guidelines&id=fc_2875171 (visité le 03/05/2023).
- [86] HEKA TEAM (INRIA-INSERM). *MedKit Library*. <https://medkit.readthedocs.io/en/stable/index.html>. Accessed : 2024-06-03. 2024.
- [87] HEMONC.ORG. *HemOnc.org : Hematology/Oncology Wiki*. https://hemonc.org/wiki/Main_Page. Accessed : 2024-06-10. 2024.
- [88] HIRONSAN. *Doccoano : Text Annotation Tool*. Accessed : 2024-08-02. 2018. URL : <https://github.com/doccano/doccano>.
- [89] Aidan HOGAN et al. « Knowledge graphs ». In : *ACM Computing Surveys (Csur)* 54.4 (2021), p. 1-37.
- [90] David W HOSMER JR, Stanley LEMESHOW et Rodney X STURDIVANT. *Applied logistic regression*. John Wiley & Sons, 2013.
- [91] WM HRYNIUK. « The importance of dose intensity in the outcome of chemotherapy. » In : *Important Adv Oncol* (1988), p. 121-141.
- [92] Le-Tian HUANG et al. « Clinical option of pemetrexed-based versus paclitaxel-based first-line chemotherapeutic regimens in combination with bevacizumab for advanced non-squamous non-small-cell lung cancer and optimal maintenance therapy : evidence from a meta-analysis of randomized control trials ». In : *BMC cancer* 21 (2021), p. 1-11.
- [93] Yan-qi HUANG et al. « Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer ». In : *Journal of clinical oncology* 34.18 (2016), p. 2157-2164.

- [94] i2B2. *Informatics for Integrating Biology & the Bedside (i2b2)*. <https://www.i2b2.org/>. Accessed : 2024-06-10. 2024.
- [95] Alexia IASONOS et al. « How to build and interpret a nomogram for cancer prognosis ». In : *Journal of clinical oncology* 26.8 (2008), p. 1364-1370.
- [96] Joseph G IBRAHIM, Haitao CHU et Liddy M CHEN. « Basic concepts and methods for joint models of longitudinal and survival data ». In : *Journal of clinical oncology* 28.16 (2010), p. 2796-2801.
- [97] INCA. *Qu'est-ce que la chimiothérapie?* Accessed : 2023. URL : <https://www.e-cancer.fr/Patients-et-proches/Se-faire-soigner/Traitements/Chimiotherapie/Qu-est-ce-que-la-chimiotherapie> (visité le 03/05/2023).
- [98] The Drip IV INFUSION. *What Is IV Bolus?* Accessed : 2024-05-26. 2024. URL : <https://thedripivinfusion.com/blog/what-is-iv-bolus>.
- [99] INSTITUT NATIONAL DU CANCER. *Panorama des cancers en France*, édition 2023. <https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Panorama-des-cancers-en-France-edition-2023>. Accédé le : 2023-05-03. 2023.
- [100] INSTITUT NATIONAL DU CANCER. *Recommandations et référentiels*. Accessed : May 3, 2023. 2023. URL : <https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Collections/Recommandations-et-referentiels> (visité le 03/05/2023).
- [101] Hemant ISHWARAN et al. « Random survival forests ». In : *The Annals of Applied Statistics* (2008).
- [102] Anne-Sophie JANNOT et al. « The Georges Pompidou University hospital clinical data warehouse : a 8-years follow-up experience ». In : *International journal of medical informatics* 102 (2017), p. 21-28.
- [103] Lamy JEAN-BAPTISTE. « Ontologies with python ». In : *Apress, Berkeley, CA* (2021).
- [104] Michael JEFFORD et al. « Improved models of care for cancer survivors ». In : *The Lancet* 399.10334 (2022), p. 1551-1560.
- [105] Jong Ho JHEE et al. « Representation and comparison of chemotherapy protocols with ChemoKG and graph embeddings ». In : *SWAT4HCLS*. 2024.
- [106] Tongchao JIANG et al. « Prognostic significance of hemoglobin, albumin, lymphocyte, and platelet (HALP) score in breast cancer : a propensity score-matching study ». In : *Cancer Cell International* 24.1 (2024), p. 230.
- [107] Fairooz KABBINAVAR et al. « Phase II, randomized trial comparing bevacizumab plus fluorouracil (FU)/leucovorin (LV) with FU/LV alone in patients with metastatic colorectal cancer ». In : *Journal of clinical oncology* 21.1 (2003), p. 60-65.

- [108] Edward L KAPLAN et Paul MEIER. « Nonparametric estimation from incomplete observations ». In : *Journal of the American statistical association* 53.282 (1958), p. 457-481.
- [109] Thu-Lan KELLY, Amy SALTER et Nicole L PRATT. « The weighted cumulative exposure method and its application to pharmacoepidemiology : A narrative review ». In : *Pharmacoepidemiology and Drug Safety* 33.1 (2024), e5701.
- [110] *Le Dictionnaire Vidal*. 98^e édition. <https://www.vidal.fr>. Paris, France : Vidal, 2023.
- [111] Charlotte LEDUC et al. « Comorbidities in the management of patients with lung cancer ». In : *European Respiratory Journal* 49.3 (2017).
- [112] Ivan LERNER, Nicolas PARIS et Xavier TANNIER. « Terminologies augmented recurrent neural network model for clinical named entity recognition ». In : *Journal of biomedical informatics* 102 (2020), p. 103356.
- [113] Fang LI et al. « Time event ontology (TEO) : to support semantic representation and reasoning of complex temporal relations of clinical events ». In : *Journal of the American Medical Informatics Association* 27.7 (2020), p. 1046-1056.
- [114] Linrong Li et al. « Development and validation of a model and nomogram for breast cancer diagnosis based on quantitative analysis of serum disease-specific haptoglobin N-glycosylation ». In : *Journal of Translational Medicine* 22.1 (2024), p. 331.
- [115] Frank P LIN et al. « Cancer care treatment outcome ontology : a novel computable ontology for profiling treatment outcomes in patients with solid tumors ». In : *JCO clinical cancer informatics* 2 (2018), p. 1-14.
- [116] Jiaxuan LIU et al. « Development and validation of a combined nomogram for predicting perineural invasion status in rectal cancer via computed tomography-based radiomics ». In : *Journal of Cancer Research and Therapeutics* 19.6 (2023), p. 1552-1559.
- [117] Sariah LIU et Razelle KURZROCK. « Toxicity of targeted therapy : Implications for response and impact of genetic polymorphisms ». In : *Cancer treatment reviews* 40.7 (2014), p. 883-891.
- [118] Zhanna LIVSHITS, Rama B RAO et Silas W SMITH. « An approach to chemotherapy-associated toxicity ». In : *Emergency Medicine Clinics* 32.1 (2014), p. 167-203.
- [119] LOINC. *LOINC : Logical Observation Identifiers Names and Codes*. <https://loinc.org/>. Accessed : 2024-06-10. 2024.
- [120] Dan L LONGO et al. « The calculation of actual or received dose intensity : a comparison of published methods ». In : *J Clin Oncol* 9.11 (1991), p. 2042-2051.
- [121] Gary H LYMAN. « Impact of chemotherapy dose intensity on cancer patient outcomes ». In : *Journal of the National Comprehensive Cancer Network* 7.1 (2009), p. 99-108.

- [122] Michael L MAITLAND, Kaveeta VASISHT et Mark J RATAIN. « TPMT, UGT1A1 and DPYD : genotyping to ensure safer cancer therapy ? » In : *Trends in pharmacological sciences* 27.8 (2006), p. 432-437.
- [123] Christopher D. MANNING et al. « The Stanford CoreNLP Natural Language Processing Toolkit ». In : *Association for Computational Linguistics (ACL) System Demonstrations*. 2014, p. 55-60. URL : <http://www.aclweb.org/anthology/P/P14/P14-5010>.
- [124] Nathan MANTEL et William HAENSZEL. « Statistical aspects of the analysis of data from retrospective studies of disease ». In : *Journal of the national cancer institute* 22.4 (1959), p. 719-748.
- [125] Nathan MANTEL et al. « Evaluation of survival data and two new rank order statistics arising in its consideration ». In : *Cancer Chemother Rep* 50.3 (1966), p. 163-170.
- [126] D MARIE et al. « Représentations sociales du cancer et de la chimiothérapie : enjeux pour la définition de la situation thérapeutique ». In : *Bulletin du cancer* 97.5 (2010), p. 577-587.
- [127] Michele MASUCCI et al. « Bridging the Divide : A Review on the Implementation of Personalized Cancer Medicine ». In : *Journal of Personalized Medicine* 14.6 (2024), p. 561.
- [128] Andrea MAURICHI et al. « Factors affecting sentinel node metastasis in thin (T1) cutaneous melanomas : development and external validation of a predictive nomogram ». In : *Journal of Clinical Oncology* 38.14 (2020), p. 1591-1601.
- [129] MEDDRA. *MedDRA*. Accessed : 2024-06-17. 2024. URL : <https://www.meddra.org/>.
- [130] Arizona IV MEDICS. *What You Need to Know About IV Bolus Administration*. Accessed : 2024-05-26. 2024. URL : <https://www.azivmedics.com/blog/what-you-need-to-know-about-iv-bolus-administration>.
- [131] Qianhao MENG et al. « Survival comparison of first-line treatment regimens in patients with braf-mutated advanced colorectal cancer : a multicenter retrospective study ». In : *BMC cancer* 23.1 (2023), p. 191.
- [132] MESH. *Medical Subject Headings (MeSH)*. <https://www.nlm.nih.gov/mesh/meshhome.html>. Accessed : 2024-06-10. 2024.
- [133] Pierre MONNIN. « Matching and mining in knowledge graphs of the Web of data-Applications in pharmacogenomics ». Thèse de doct. Université de Lorraine, 2020.
- [134] Pierre MONNIN et al. « Discovering alignment relations with Graph Convolutional Networks : A biomedical case study ». In : *Semantic Web* 13.3 (2022), p. 379-398.
- [135] Pierre MONNIN et al. « PGxO and PGxLOD : a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison ». In : *BMC Bioinformatics* 20-S.4 (2019), p. 139. ISSN : 1471-2105. DOI : [10.1186/s12859-019-2693-9](https://doi.org/10.1186/s12859-019-2693-9). URL : <https://doi.org/10.1186/s12859-019-2693-9>.

- [136] Boris MOTIK, Rob SHEARER et Ian HORROCKS. « HermiT : A Highly-Efficient OWL Reasoner ». In : *Proceedings of the 5th International Workshop on OWL : Experiences and Directions (OWLED 2008)*. 2008. URL : <http://www.hermit-reasoner.com/>.
- [137] Juliette MURRIS. « Les forêts aléatoires de survie pour l'analyse des événements récurrents ». Thèse en cours d'écriture. 2024.
- [138] Yousuke NAKAI et al. « Comorbidity, not age, is prognostic in patients with advanced pancreatic cancer receiving gemcitabine-based chemotherapy ». In : *Critical reviews in oncology/hematology* 78.3 (2011), p. 252-259.
- [139] NATIONAL CANCER INSTITUTE. *NCI Thesaurus (NCIt) Browser*. https://ncit.nci.nih.gov/ncitbrowser/pages/multiple_search.jsf?nav_type=terminologies. Accessed : 2024-06-10. 2024.
- [140] NATIONAL CANCER INSTITUTE. *What Is Cancer?* Accessed : 2024-08-02. 2024. URL : <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>.
- [141] NATIONAL COMPREHENSIVE CANCER NETWORK. *NCCN Clinical Practice Guidelines in Oncology*. Accessed : 2023. URL : <https://www.nccn.org/guidelines/nccn-guidelines> (visité le 03/05/2023).
- [142] NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE). *Guidance by Cancer Type*. Accessed : May 3, 2023. 2023. URL : <https://www.nice.org.uk/guidance/conditions-and-diseases/cancer/> (visité le 03/05/2023).
- [143] Gonzalo NAVARRO. « A guided tour to approximate string matching ». In : *ACM computing surveys (CSUR)* 33.1 (2001), p. 31-88.
- [144] Saul B NEEDLEMAN et Christian D WUNSCH. « A general method applicable to the search for similarities in the amino acid sequence of two proteins ». In : *Journal of molecular biology* 48.3 (1970), p. 443-453.
- [145] Antoine NEURAZ et al. « Natural language processing for rapid response to emergent diseases : case study of calcium channel blockers and hypertension in the COVID-19 pandemic ». In : *Journal of medical Internet research* 22.8 (2020), e20773.
- [146] Carrie M NIELSON et al. « Relative dose intensity of chemotherapy and survival in patients with advanced stage solid tumor cancer : a systematic review and meta-analysis ». In : *The Oncologist* 26.9 (2021), e1609-e1618.
- [147] Ginah NIGHTINGALE et al. « Clinical pharmacology of oncology agents in older adults : a comprehensive review of how chronologic and functional age can influence treatment-related effects ». In : *Journal of geriatric oncology* 10.1 (2019), p. 4-30.

- [148] MA NOORBHAI et al. « Elevated international normalised ratios correlate with severity of injury and outcome ». In : *South African Medical Journal* 106.11 (2016), p. 1141-1145.
- [149] Mehdi NOURELAHI et al. « A model to predict breast cancer survivability using logistic regression ». In : *Middle East Journal of Cancer* 10.2 (2019), p. 132-138.
- [150] OBI ONTOLOGY. *Ontology for Biomedical Investigations*. <https://obi-ontology.org/>. Accessed : 2024-06-10. 2024.
- [151] OHDSI. *OHDSI : Observational Health Data Sciences and Informatics*. <https://www.ohdsi.org/data-standardization/>. Accessed : 2024-06-10. 2024.
- [152] Naoaki OKAZAKI et Jun'ichi TSUJII. « Simple and efficient algorithm for approximate dictionary matching ». In : *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010, p. 851-859.
- [153] M. M. OKEN et al. « Toxicity And Response Criteria Of The Eastern Cooperative Oncology Group ». In : *American Journal of Clinical Oncology* 5 (6 1982), p. 649-655. URL : <https://ecog-acrin.org/resources/ecog-performance-status/>.
- [154] ONTOTEXT. *GraphDB : The RDF Database*. Version 9.11, <https://www.ontotext.com/products/graphdb/>. 2023.
- [155] World Health ORGANIZATION et al. *Adherence to long-term therapies : evidence for action*. World Health Organization, 2003.
- [156] AR PADHANI et L OLLIVIER. « The RECIST criteria : implications for diagnostic radiologists ». In : *The British journal of radiology* 74.887 (2001), p. 983-986.
- [157] Vinita B PAI et Milap C NAHATA. « Cardiotoxicity of chemotherapeutic agents : incidence, treatment and prevention ». In : *Drug safety* 22 (2000), p. 263-302.
- [158] PGxLOD. *PGxLOD : Pharmacogenomics Linked Open Data*. <https://pgxlod.loria.fr/>. Accessed : 2024-06-10. 2024.
- [159] Jane Louise PHILLIPS et David C CURROW. « Cancer as a chronic disease ». In : *Collegian* 17.2 (2010), p. 47-50.
- [160] P PIEDBOIS et al. « Efficacy of intravenous continuous infusion of fluorouracil compared with bolus administration in advanced colorectal cancer. » In : *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 16.1 (1998), p. 301-308.
- [161] Clémence POIROT. « L'information sur les effets indésirables de la chimiothérapie anticancéreuse : les besoins du patient et la place du pharmacien ». Thèse de doct. Thèse d'exercice Pharmacie. Lorraine :[sn], 2014.
- [162] Boston University School of PUBLIC HEALTH. *Survival Analysis*. Accessed : 2024-08-02. n.d. URL : https://sphweb.bumc.bu.edu/otlt MPH-Modules/BS/BS704_Survival/index.html#headingtaglink_1.

- [163] Peng QI et al. « Stanza : A Python natural language processing toolkit for many human languages ». In : *arXiv preprint arXiv :2003.07082* (2020).
- [164] Elisabeth QUOIX et al. « Carboplatin and weekly paclitaxel doublet chemotherapy compared with monotherapy in elderly patients with advanced non-small-cell lung cancer : IFCT-0501 randomised, phase 3 trial ». In : *The Lancet* 378.9796 (2011), p. 1079-1088.
- [165] Mark J RATAIN, Daniel A GOLDSTEIN et Allen S LICHTER. « Interventional pharmacoeconomics—a new discipline for a cost-constrained environment ». In : *JAMA oncology* 5.8 (2019), p. 1097-1098.
- [166] European Organisation for RESEARCH et Treatment of CANCER. *EORTC Quality of Life Questionnaires*. Accessed : 2024-07-23. n.d. URL : <https://qol.eortc.org/questionnaires/>.
- [167] Jacques ROBERT et al. « Predicting drug response and toxicity based on gene polymorphisms ». In : *Critical reviews in oncology/hematology* 54.3 (2005), p. 171-196.
- [168] George RODRIGUES et Michael SANATANI. « Age and comorbidity considerations related to radiotherapy and chemotherapy administration ». In : *Seminars in radiation oncology*. T. 22. 4. Elsevier. 2012, p. 277-283.
- [169] Alice ROGIER, Adrien COULET et Bastien RANCE. « Using an ontological representation of chemotherapy toxicities for guiding information extraction and integration from EHRs ». In : *MEDINFO 2021 : One World, One Health–Global Partnership for Digital Innovation*. IOS Press, 2022, p. 91-95.
- [170] Alice ROGIER, Bastien RANCE et Adrien COULET. *ChemoOnto, an ontology to qualify the course of chemotherapies*. Jan. 2024. doi : 10.5281/zenodo.10548491. URL : <https://doi.org/10.5281/zenodo.10548491>.
- [171] ROMEDI. *ROMEDI Ontology*. <https://bioportal.lirmm.fr/ontologies/ROMEDI/?p=classes&conceptid=root>. Accessed : 2024-06-10. 2024.
- [172] Ryan D ROSEN et Amit SAPRA. « TNM classification ». In : *StatPearls [Internet]*. StatPearls Publishing, 2023.
- [173] Samuel M RUBINSTEIN et al. « Standardizing chemotherapy regimen nomenclature : a proposal and evaluation of the HemOnc and National Cancer Institute Thesaurus Regimen Content ». In : *JCO Clinical Cancer Informatics* 4 (2020), p. 60-70.
- [174] RxNORM. *RxNorm*. <https://www.nlm.nih.gov/research/umls/rxnorm/index.html>. Accessed : 2024-06-10. 2024.
- [175] Jeffrey R SACHS et al. « Optimal dosing for targeted therapies in oncology : drug development cases leading by example ». In : *Clinical Cancer Research* 22.6 (2016), p. 1318-1324.

- [176] Leonard B SALTZ et al. « Bevacizumab in combination with oxaliplatin-based chemotherapy as first-line therapy in metastatic colorectal cancer : a randomized phase III study ». In : *Journal of clinical oncology* 26.12 (2008), p. 2013-2019.
- [177] Sagar SARDESAI et al. « Clinical impact of interruption in adjuvant Trastuzumab therapy in patients with operable HER-2 positive breast cancer ». In : *Cardio-Oncology* 6 (2020), p. 1-9.
- [178] C Coscarelli SCHAG, Richard L HEINRICH et Patricia A GANZ. « Karnofsky performance status revisited : reliability, validity, and guidelines. » In : *Journal of Clinical Oncology* 2.3 (1984), p. 187-193.
- [179] David SCHOENFELD. « Partial residuals for the proportional hazards regression model ». In : *Biometrika* 69.1 (1982), p. 239-241.
- [180] Guus SCHREIBER. « Knowledge engineering ». In : *Foundations of Artificial Intelligence* 3 (2008), p. 929-946.
- [181] Guus SCHREIBER. *Knowledge engineering and management : the CommonKADS methodology*. MIT press, 2000.
- [182] Erich SCHUBERT. « A triangle inequality for cosine similarity ». In : *International Conference on Similarity Search and Applications*. Springer. 2021, p. 32-44.
- [183] Claudia SEW SCHUURHUIZEN et al. « The predictive value of cumulative toxicity for quality of life in patients with metastatic colorectal cancer during first-line palliative chemotherapy ». In : *Cancer management and research* (2018), p. 3015-3021.
- [184] Paul SEIFERT et al. « Comparison of continuously infused 5-fluorouracil with bolus injection in treatment of patients with colorectal adenocarcinoma ». In : *Cancer* 36.1 (1975), p. 123-128.
- [185] Peter H SELLERS. « On the theory and computation of evolutionary distances ». In : *SIAM Journal on Applied Mathematics* 26.4 (1974), p. 787-793.
- [186] Nigam H SHAH et al. « Proton pump inhibitor usage and the risk of myocardial infarction in the general population ». In : *PLoS one* 10.6 (2015), e0124653.
- [187] Evren SIRIN et al. « Pellet : A Practical OWL-DL Reasoner ». In : *Web Semantics : Science, Services and Agents on the World Wide Web*. T. 5. 2. 2007, p. 51-53. URL : <https://github.com/Complexible/pellet>.
- [188] Temple F SMITH, Michael S WATERMAN et al. « Identification of common molecular subsequences ». In : *Journal of molecular biology* 147.1 (1981), p. 195-197.
- [189] SNOMED INTERNATIONAL. *SNOMED CT : Systematized Nomenclature of Medicine – Clinical Terms*. <https://www.snomed.org/>. Accessed : 2024-06-10. 2024.
- [190] OpenLink SOFTWARE. *Virtuoso Universal Server*. Version 7.2.7, <https://virtuoso.openlinksw.com/>. 2023.
- [191] Mette SØGAARD et al. « The impact of comorbidity on cancer survival : a review ». In : *Clinical epidemiology* 5.sup1 (2013), p. 3-29.

- [192] Luca SOLDAINI et Nazli GOHARIAN. « Quickumls : a fast, unsupervised approach for medical concept extraction ». In : *MedIR workshop, sigir*. 2016, p. 1-4.
- [193] Martin STANULLA et al. « Thiopurine methyltransferase (TPMT) genotype and early treatment response to mercaptopurine in childhood acute lymphoblastic leukemia ». In : *Jama* 293.12 (2005), p. 1485-1489.
- [194] Pontus STENETORP et al. « brat : a Web-based Tool for NLP-Assisted Text Annotation ». In : *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Accessed : 2024-08-02. 2012, p. 102-107. URL : <https://brat.nlplab.org/>.
- [195] Krithika SURESH, Cameron SEVERN et Debasish GHOSH. « Survival prediction models : an introduction to discrete-time modeling ». In : *BMC medical research methodology* 22.1 (2022), p. 207.
- [196] Masashi TAKANO et Toru SUGIYAMA. « UGT1A1 polymorphisms in cancer : impact on irinotecan treatment ». In : *Pharmacogenomics and personalized medicine* (2017), p. 61-68.
- [197] Andrew TANG et al. « How much delay matters ? How time to treatment impacts overall survival in early stage lung cancer ». In : *Annals of Surgery* (2022).
- [198] Cui TAO, Harold R SOLBRIG et Christopher G CHUTE. « CNTRON 2.0 : a harmonized semantic web ontology for temporal relation inferencing in clinical narratives ». In : *AMIA summits on translational science proceedings* 2011 (2011), p. 64.
- [199] Cui TAO et al. « CNTRON : a semantic web ontology for temporal relation inferencing in clinical narratives ». In : *AMIA annual symposium proceedings*. T. 2010. American Medical Informatics Association. 2010, p. 787.
- [200] Nicholas P TATONETTI et al. « Detecting drug interactions from adverse-event reports : interaction between paroxetine and pravastatin increases blood glucose levels ». In : *Clinical Pharmacology & Therapeutics* 90.1 (2011), p. 133-142.
- [201] David THOMPSON. *Replication of randomized, controlled trials using real-world data : what could go wrong ?* 2021.
- [202] Robert TIBSHIRANI. « Regression shrinkage and selection via the lasso ». In : *Journal of the Royal Statistical Society Series B : Statistical Methodology* 58.1 (1996), p. 267-288.
- [203] Elizabeth TOLL. « The cost of technology ». In : *Jama* 307.23 (2012), p. 2497-2498.
- [204] Matthew R TRENDOWSKI et al. « Genetic and modifiable risk factors contributing to cisplatin-induced toxicities ». In : *Clinical Cancer Research* 25.4 (2019), p. 1147-1155.
- [205] Hajime UNO et al. « On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data ». In : *Statistics in medicine* 30.10 (2011), p. 1105-1117.
- [206] U.S. NATIONAL LIBRARY OF MEDICINE. *Unified Medical Language System (UMLS)*. <https://www.nlm.nih.gov/research/umls/index.html>. Accessed : 2024-06-10. 2024.

- [207] Pranjal VAIDYA et al. « CT derived radiomic score for predicting the added benefit of adjuvant chemotherapy following surgery in stage I, II resectable non-small cell lung cancer : a retrospective multicohort study for outcome prediction ». In : *The Lancet Digital Health* 2.3 (2020), e116-e128.
- [208] André BP VAN KUILENBURG. « Dihydropyrimidine dehydrogenase and the efficacy and toxicity of 5-fluorouracil ». In : *European journal of cancer* 40.7 (2004), p. 939-950.
- [209] W3C OWL WORKING GROUP. *OWL 2 Web Ontology Language Primer (Second Edition)*. <https://www.w3.org/TR/owl2-primer/>. [Online; accessed 5-June-2024]. 2012.
- [210] W3C RDF WORKING GROUP. *RDF Schema 1.1*. <https://www.w3.org/TR/2014/REC-rdf-schema-20140225/>. [Online; accessed 5-June-2024]. 2014.
- [211] Tian Qi WANG et al. « Routine surveillance of chemotherapy toxicities in cancer patients using the patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE) ». In : *Oncology and Therapy* 6 (2018), p. 189-201.
- [212] Jeremy L WARNER et al. « HemOnc : A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model ». In : *Journal of biomedical informatics* 96 (2019), p. 103239.
- [213] Jeremy L WARNER et al. « HemOnc.org : A collaborative online knowledge platform for oncology professionals ». In : *Journal of Oncology Practice* 11.3 (2015), e336-e350.
- [214] Qing WEI et al. « The promise and challenges of combination therapies with antibody-drug conjugates in solid tumors ». In : *Journal of Hematology & Oncology* 17.1 (2024), p. 1.
- [215] WHO. *WHO model list of essential medicines–22nd list*. 2021.
- [216] WHO-UMC. *WHO-ART : WHO Adverse Reactions Terminology*. <https://who-umc.org/vigibase/vigibase-services/who-art/>. Accessed : 2024-06-10. 2024.
- [217] WIKIPÉDIA. *Distance (mathématiques)*. [https://fr.wikipedia.org/wiki/Distance_\(math%C3%A9matiques\)](https://fr.wikipedia.org/wiki/Distance_(math%C3%A9matiques)). Accessed : 2024-08-02. 2024.
- [218] WIKIPÉDIA. *Mesure (mathématiques)*. [https://fr.wikipedia.org/wiki/Mesure_\(math%C3%A9matiques\)](https://fr.wikipedia.org/wiki/Mesure_(math%C3%A9matiques)). Accessed : 2024-08-02. 2024.
- [219] WORLD HEALTH ORGANIZATION. *ATC/DDD Toolkit : ATC Classification*. <https://www.who.int/tools/atc-ddd-toolkit/atc-classification>. Accessed : 2024-06-10. 2024.
- [220] WORLD HEALTH ORGANIZATION. *International Classification of Diseases (ICD-10)*. <https://icd.who.int/browse/2024-01/mms/fr>. Accessed : 2024-06-10. 2024.

- [221] Victoria S Wu et al. « Why the treatment sequence matters : interplay between chemotherapy cycles received, cumulative dose intensity, and survival in resected early-stage pancreas cancer ». In : *Annals of surgery* 278.4 (2023), e677-e684.
- [222] Gwen WYATT et al. « Chemotherapy interruptions in relation to symptom severity in advanced breast cancer ». In : *Supportive Care in Cancer* 23 (2015), p. 3183-3191.
- [223] Emily C ZABOR et al. « Logistic regression in clinical studies ». In : *International Journal of Radiation Oncology* Biology* Physics* 112.2 (2022), p. 271-277.
- [224] Eric ZAPLETAL et al. « Methodology of integration of a clinical data warehouse with a clinical information system : the HEGP case ». In : *MEDINFO 2010*. IOS Press, 2010, p. 193-197.
- [225] Charles G ZUBROD et al. « Appraisal of methods for the study of chemotherapy of cancer in man : comparative therapeutic trial of nitrogen mustard and triethylene thiophosphoramide ». In : *Journal of Chronic Diseases* 11.1 (1960), p. 7-33.