Année 2020/2021 N°

Thèse

Pour le

DOCTORAT EN MÉDECINE

Diplôme d'État

par

Marie ANSOBORLO

Née le 2 avril 1992 à Bordeaux (33)

OUTIL D'AIDE A LA PRÉ-SÉLECTION DANS LES ESSAIS CLINIQUES EN PNEUMO-ONCOLOGIE : APPARIER LES FICHES RCP ET LES PROTOCOLES DU REGISTRE FRANÇAIS.

Présentée et soutenue publiquement le **20 octobre 2021** devant un jury composé de :

Président du Jury : Professeur Emmanuel RUSCH, Epidémiologie, économie de la santé et prévention, Faculté de Médecine – Tours

Membres du Jury :
Professeur Claude LINASSIER, Cancérologie, radiothérapie, Faculté de Médecine – Tours
Docteur Chloé PLICHON, Biopharmacie clinique oncologique, PH, CHU – Tours

Directeur de thèse : Docteur Leslie GUILLON–GRAMMATICO, Epidémiologie, économie de la santé et prévention, MCU-PH, HDR, Faculté de Médecine – Tours

Résumé

Frein majeur à l'inclusion dans les essais cliniques, la pré-sélection des sujets éligibles par la fouille manuelle des dossiers des patients est chronophage, mais peut être accélérée grâce à l'exploitation des dossiers informatisés. Augmenter l'efficacité et l'exhaustivité de l'identification des patients potentiellement éligibles à l'entrée dans les essais en pneumo-oncologie est possible en développant un outil de traitement automatique du langage pour vérifier les critères d'inclusion.

Pour chacun des huit critères d'inclusion aux essais cliniques étudiés, des algorithmes basés sur des expressions régulières ont été implémentés et évaluées. L'appariement sujet-protocole a été estimé entre les comptes rendus de Réunion de Concertation Pluri-professionnelles (RCP) stockés dans l'entrepôt de données hospitalières du CHRU de Tours et les protocoles publiés dans le registre français de l'institut contre le cancer (RECF) concernant les essais thérapeutiques ouverts à l'inclusion dans le Grand Ouest.

Ont été extraits 368 comptes rendus de RCP entre janvier et juillet 2021 ainsi que 16 protocoles d'essais ouverts en juillet 2021. Les performances pour détecter les critères d'inclusion dans les essais cliniques s'élevaient à 86 % de précision et 89 % de rappel en moyenne par protocole. L'outil appariait au moins un protocole à près d'un tiers des sujets (29,6 %) et identifiait au moins un sujet éligible pour la majorité des protocoles (81,3 %).

La pré-sélection automatisée des patients de pneumo-oncologie diminuerait la charge de travail avec une performance au moins aussi élevée que celle retrouvée dans la littérature. Les comptes rendus de RCP étant plus structurés que les fiches des protocoles, l'extraction des caractéristiques du patient montrait des performances plus élevées que celle des critères d'éligibilité. La précision de l'appariement sujet-protocole pourrait être augmentée en utilisant des variables de biologie. Tester l'outil sur les comptes rendus de RCP d'autres centres serait nécessaire pour estimer sa validité externe et permettre une interopérabilité des entrepôts de données pour des recrutements multicentriques, notamment avec la disponibilité du Ouest Data Hub.

Mots clés

"Tumeurs pulmonaires/statistiques et données numériques",

"Stades des tumeurs/utilisation thérapeutique",

"Réunion de concertation multidisciplinaire",

"Traitement automatique du langage naturel",

"Pré-sélection du traitement du patient".

Title

“Automatic trial eligibility on lung cancer based on natural language processing”.

Abstract

A major issue for inclusion in clinical trials (CTs) is the time-consuming chart review process, which can be accelerated by natural language processing. To increase the efficiency and completeness of the patients pre-screening for lung cancer CT, the objective was to develop and evaluate a matching tool between the Multi-Professional Consultation Meeting (MCM) reports and the CT registered by the French National Cancer Institute (INCa).

Algorithms based on regular expressions were evaluated for 7 inclusion criteria related to the patient-cancer profile. Matching between the MCM reports from January to July 2021, stored in the hospital data warehouse of the CHRU of Tours, and the CTs open to inclusion in the Western region in July 2021 in the French Cancer Institute Registry (RECF) was estimated.

From the data warehouse, 368 MCM were extracted between January and July 2021 and 16 CT protocols opened in July 2021 in the RECF. The performance for information extraction from the protocols reached 86% of precision and 89% of recall in macroaverage. The tool proposed one protocol or more at one quarter patients (29.6%) and identified eligible patients for the majority of the protocols (81.3%).

The tool performed better on patient-MCM reports than on RECF trials extracting eligibility criteria. To automate the prescreening of lung cancer patients for entering a CT could highly reduce the workload of healthcare and research professionals with a good performance. The accuracy could be increased by using biological variables. Testing the tool on other oncology centers having data warehouse is necessary to assess its external validity, especially with the availability of the West Data Hub.

Key Words

"Lung Neoplasms/statistics and numerical data",

"Neoplasm Staging/therapeutic use",

“Multidisciplinary team meeting consultation”,

“Natural language processing”,

“Patient Treatment prescreening”.

UNIVERSITE DE TOURS
FACULTE DE MEDECINE DE TOURS

DOYEN
Pr Patrice DIOT

VICE-DOYEN
Pr Henri MARRET

ASSESSEURS
Pr Denis ANGOULVANT, *Pédagogie*
Pr Mathias BUCHLER, *Relations internationales*
Pr Theodora BEJAN-ANGOULVANT, *Moyens – relations avec l'Université*
Pr Clarisse DIBAO-DINA, *Médecine générale*
Pr François MAILLOT, *Formation Médicale Continue*
Pr Patrick VOURC'H, *Recherche*

RESPONSABLE ADMINISTRATIVE
Mme Fanny BOBLETER

****

DOYENS HONORAIRES
Pr Emile ARON (†) – *1962-1966*
Directeur de l'Ecole de Médecine - 1947-1962
Pr Georges DESBUQUOIS (†) - *1966-1972*
Pr André GOUAZE (†) - *1972-1994*
Pr Jean-Claude ROLLAND – *1994-2004*
Pr Dominique PERROTIN – *2004-2014*

PROFESSEURS EMERITES
Pr Daniel ALISON
Pr Gilles BODY
Pr Jacques CHANDENIER
Pr Philippe COLOMBAT
Pr Etienne DANQUECHIN-DORVAL
Pr Pascal DUMONT
Pr Dominique GOGA
Pr Gérard LORETTE
Pr Dominique PERROTIN
Pr Roland QUENTIN

PROFESSEURS HONORAIRES
P. ANTHONIOZ – P. ARBEILLE – A. AUDURIER – A. AUTRET – P. BAGROS – P.BARDOS – C. BARTHELEMY – J.L. BAULIEU – C. BERGER – JC. BESNARD – P. BEUTTER – C. BONNARD – P. BONNET – P. BOUGNOUX – P. BURDIN – L. CASTELLANI – A. CHANTEPIE – B. CHARBONNIER – P. CHOUTET – T. CONSTANS – P. COSNAY – C. COUET – L. DE LA LANDE DE CALAN – J.P. FAUCHIER – F. FETISSOF – J. FUSCIARDI – P. GAILLARD – G. GINIES – A. GOUDEAU – J.L. GUILMOT – O. HAILLOT – N. HUTEN – M. JAN – J.P. LAMAGNERE – F. LAMISSE – Y. LANSON – O. LE FLOCH – Y. LEBRANCHU – E. LECA – P. LECOMTE – AM. LEHR-DRYLEWICZ – E. LEMARIE – G. LEROY – M. MARCHAND – C. MAURAGE – C. MERCIER – J. MOLINE – C. MORAINE – J.P. MUH – J. MURAT – H. NIVET – L. POURCELOT – P. RAYNAUD – D. RICHARD-LENOBLE – A. ROBIER – J.C. ROLLAND – D. ROYERE - A. SAINDELLE – E. SALIBA – J.J. SANTINI – D. SAUVAGE – D. SIRINELLI – J. WEILL

PROFESSEURS DES UNIVERSITES - PRATICIENS HOSPITALIERS

ANDRES Christian.....Biochimie et biologie moléculaire
ANGOULVANT Denis .....Cardiologie
APETOH Lionel .....Immunologie
AUPART Michel.....Chirurgie thoracique et cardiovasculaire
BABUTY Dominique .....Cardiologie
BAKHOS David .....Oto-rhino-laryngologie
BALLON Nicolas .....Psychiatrie ; addictologie
BARILLOT Isabelle .....Cancérologie ; radiothérapie
BARON Christophe .....Immunologie
BEJAN-ANGOULVANT Théodora .....Pharmacologie clinique
BERHOUET Julien .....Chirurgie orthopédique et traumatologique
BERNARD Anne .....Cardiologie
BERNARD Louis .....Maladies infectieuses et maladies tropicales
BLANCHARD-LAUMONNIER Emmanuelle ....Biologie cellulaire
BLASCO Hélène.....Biochimie et biologie moléculaire
BONNET-BRILHAULT Frédérique .....Physiologie
BOURGUIGNON Thierry .....Chirurgie thoracique et cardiovasculaire
BRILHAULT Jean.....Chirurgie orthopédique et traumatologique
BRUNEREAU Laurent .....Radiologie et imagerie médicale
BRUYERE Franck.....Urologie
BUCHLER Matthias.....Néphrologie
CALAIS Gilles.....Cancérologie, radiothérapie
CAMUS Vincent .....Psychiatrie d'adultes
CORCIA Philippe.....Neurologie
COTTIER Jean-Philippe .....Radiologie et imagerie médicale
DE TOFFOL Bertrand .....Neurologie
DEQUIN Pierre-François.....Thérapeutique
DESOUBEAUX Guillaume.....Parasitologie et mycologie
DESTRIEUX Christophe .....Anatomie
DIOT Patrice.....Pneumologie
DU BOUEXIC de PINIEUX Gonzague .....Anatomie & cytologie pathologiques
DUCLUZEAU Pierre-Henri .....Endocrinologie, diabétologie, et nutrition
EL HAGE Wissam.....Psychiatrie adultes
EHRMANN Stephan .....Médecine intensive – réanimation
FAUCHIER Laurent .....Cardiologie
FAVARD Luc.....Chirurgie orthopédique et traumatologique
FOUGERE Bertrand .....Gériatrie
FOUQUET Bernard.....Médecine physique et de réadaptation
FRANCOIS Patrick.....Neurochirurgie
FROMONT-HANKARD Gaëlle .....Anatomie & cytologie pathologiques
GATAULT Philippe.....Néphrologie
GAUDY-GRAFFIN Catherine.....Bactériologie-virologie, hygiène hospitalière
GOUPILLE Philippe .....Rhumatologie
GRUEL Yves.....Hématologie, transfusion
GUERIF Fabrice .....Biologie et médecine du développement et de la reproduction
GUILLON Antoine.....Médecine intensive – réanimation
GUYETANT Serge .....Anatomie et cytologie pathologiques
GYAN Emmanuel.....Hématologie, transfusion
HALIMI Jean-Michel.....Thérapeutique
HANKARD Régis.....Pédiatrie
HERAULT Olivier .....Hématologie, transfusion
HERBRETEAU Denis .....Radiologie et imagerie médicale
HOURIOUX Christophe.....Biologie cellulaire
IVANES Fabrice .....Physiologie
LABARTHE François .....Pédiatrie
LAFFON Marc .....Anesthésiologie et réanimation chirurgicale, médecine d'urgence
LARDY Hubert.....Chirurgie infantile
LARIBI Saïd.....Médecine d'urgence
LARTIGUE Marie-Frédérique .....Bactériologie-virologie
LAURE Boris .....Chirurgie maxillo-faciale et stomatologie
LECOMTE Thierry.....Gastroentérologie, hépatologie
LESCANNE Emmanuel.....Oto-rhino-laryngologie
LINASSIER Claude .....Cancérologie, radiothérapie
MACHET Laurent .....Dermato-vénéréologie
MAILLOT François .....Médecine interne

MARCHAND-ADAM Sylvain .....Pneumologie
MARRET Henri .....Gynécologie-obstétrique
MARUANI Annabel .....Dermatologie-vénéréologie
MEREGHETTI Laurent .....Bactériologie-virologie ; hygiène hospitalière
MITANCHEZ Delphine .....Pédiatrie
MORINIERE Sylvain.....Oto-rhino-laryngologie
MOUSSATA Driffa .....Gastro-entérologie
MULLEMAN Denis.....Rhumatologie
ODENT Thierry.....Chirurgie infantile
OUAISSI Mehdi .....Chirurgie digestive
OULDAMER Lobna.....Gynécologie-obstétrique
PAINTAUD Gilles .....Pharmacologie fondamentale, pharmacologie clinique
PATAT Frédéric .....Biophysique et médecine nucléaire
PERROTIN Franck .....Gynécologie-obstétrique
PISELLA Pierre-Jean.....Ophtalmologie
PLANTIER Laurent.....Physiologie
REMERAND Francis.....Anesthésiologie et réanimation, médecine d'urgence
ROINGEARD Philippe.....Biologie cellulaire
ROSSET Philippe.....Chirurgie orthopédique et traumatologique
RUSCH Emmanuel.....Epidémiologie, économie de la santé et prévention
SAINT-MARTIN Pauline.....Médecine légale et droit de la santé
SALAME Ephrem.....Chirurgie digestive
SAMIMI Mahtab.....Dermatologie-vénéréologie
SANTIAGO-RIBEIRO Maria .....Biophysique et médecine nucléaire
THOMAS-CASTELNAU Pierre .....Pédiatrie
TOUTAIN Annick.....Génétique
VAILLANT Loïc.....Dermato-vénéréologie
VELUT Stéphane.....Anatomie
VOURC'H Patrick.....Biochimie et biologie moléculaire
WATIER Hervé .....Immunologie
ZEMMOURA Ilyess .....Neurochirurgie

PROFESSEUR DES UNIVERSITES DE MEDECINE GENERALE

DIBAO-DINA Clarisse
LEBEAU Jean-Pierre

PROFESSEURS ASSOCIES

MALLET Donatien.....Soins palliatifs
POTIER Alain.....Médecine Générale
ROBERT Jean.....Médecine Générale

PROFESSEUR CERTIFIE DU 2ND DEGRE

MC CARTHY Catherine.....Anglais

MAITRES DE CONFERENCES DES UNIVERSITES - PRATICIENS HOSPITALIERS

AUDEMARD-VERGER Alexandra.....Médecine interne
BARBIER Louise.....Chirurgie digestive
BINET Aurélien .....Chirurgie infantile
BISSON Arnaud .....Cardiologie (CHRO)
BRUNAULT Paul .....Psychiatrie d'adultes, addictologie
CAILLE Agnès .....Biostat., informatique médical et technologies de communication
CARVAJAL-ALLEGRIA Guillermo.....Rhumatologie (au 01/10/2021)
CLEMENTY Nicolas.....Cardiologie
DENIS Frédéric.....Odontologie
DOMELIER Anne-Sophie .....Bactériologie-virologie, hygiène hospitalière
DUFOUR Diane .....Biophysique et médecine nucléaire
ELKRIEF Laure.....Hépatologie – gastroentérologie
FAVRAIS Géraldine.....Pédiatrie
FOUQUET-BERGEMER Anne-Marie.....Anatomie et cytologie pathologiques
GOUILLEUX Valérie.....Immunologie

GUILLON-GRAMMATICO Leslie.....Epidémiologie, économie de la santé et prévention
HOARAU Cyrille.....Immunologie
LE GUELLEC Chantal.....Pharmacologie fondamentale, pharmacologie clinique
LEFORT Bruno.....Pédiatrie
LEGRAS Antoine.....Chirurgie thoracique
LEMAIGNEN Adrien.....Maladies infectieuses
MACHET Marie-Christine.....Anatomie et cytologie pathologiques
MOREL Baptiste.....Radiologie pédiatrique
PARE Arnaud.....Chirurgie maxillo-faciale et stomatologie
PIVER Éric.....Biochimie et biologie moléculaire
REROLLE Camille.....Médecine légale
ROUMY Jérôme.....Biophysique et médecine nucléaire
SAUTENET Bénédicte.....Thérapeutique
STANDLEY-MIQUELESTORENA Elodie.....Anatomie et cytologie pathologiques
STEFIC Karl.....Bactériologie
TERNANT David.....Pharmacologie fondamentale, pharmacologie clinique
VUILLAUME-WINTER Marie-Laure.....Génétique

MAITRES DE CONFERENCES DES UNIVERSITES

AGUILLON-HERNANDEZ Nadia.....Neurosciences
NICOGLOU Antonine.....Philosophie – histoire des sciences et des techniques
PATIENT Romuald.....Biologie cellulaire
RENOUX-JACQUET Cécile.....Médecine Générale

MAITRES DE CONFERENCES ASSOCIES

BARBEAU Ludivine.....Médecine Générale
RUIZ Christophe.....Médecine Générale
SAMKO Boris.....Médecine Générale

CHERCHEURS INSERM - CNRS - INRAE

BECKER Jérôme.....Chargé de Recherche Inserm – UMR Inserm 1253
BOUAKAZ Ayache.....Directeur de Recherche Inserm – UMR Inserm 1253
BRIARD Benoit.....Chargé de Recherche Inserm – UMR Inserm 1100
CHALON Sylvie.....Directeur de Recherche Inserm – UMR Inserm 1253
DE ROCQUIGNY Hugues.....Chargé de Recherche Inserm – UMR Inserm 1259
ESCOFFRE Jean-Michel.....Chargé de Recherche Inserm – UMR Inserm 1253
GILOT Philippe.....Chargé de Recherche Inrae – UMR Inrae 1282
GOUILLEUX Fabrice.....Directeur de Recherche CNRS – EA 7501 - ERL CNRS 7001
GOMOT Marie.....Chargée de Recherche Inserm – UMR Inserm 1253
HEUZE-VOURCH Nathalie.....Directrice de Recherche Inserm – UMR Inserm 1100
KORKMAZ Brice.....Chargé de Recherche Inserm – UMR Inserm 1100
LATINUS Marianne.....Chargée de Recherche Inserm – UMR Inserm 1253
LAUMONNIER Frédéric.....Chargé de Recherche Inserm - UMR Inserm 1253
LE MERREUR Julie.....Directrice de Recherche CNRS – UMR Inserm 1253
MAMMANO Fabrizio.....Directeur de Recherche Inserm – UMR Inserm 1259
MEUNIER Jean-Christophe.....Chargé de Recherche Inserm – UMR Inserm 1259
PAGET Christophe.....Chargé de Recherche Inserm – UMR Inserm 1100
RAOUL William.....Chargé de Recherche Inserm – UMR CNRS 1069
SI TAHAR Mustapha.....Directeur de Recherche Inserm – UMR Inserm 1100
SUREAU Camille.....Directrice de Recherche émérite CNRS – UMR Inserm 1259
WARDAK Claire.....Chargée de Recherche Inserm – UMR Inserm 1253

CHARGES D'ENSEIGNEMENT

Pour l'Ecole d'Orthophonie
DELORE Claire.....Orthophoniste
GOUIN Jean-Marie.....Praticien Hospitalier
Pour l'Ecole d'Orthoptie
BOULNOIS Sandrine.....Orthoptiste
SALAME Najwa.....Orthoptiste
Pour l'Ethique Médicale
BIRMELE Béatrice.....Praticien Hospitalier

Serment d'Hippocrate

« En présence des Maîtres de cette Faculté,

de mes chers condisciples

et selon la tradition d'Hippocrate,

je promets et je jure d'être fidèle aux lois de l'honneur

et de la probité dans l'exercice de la Médecine.

Je donnerai mes soins gratuits à l'indigent,

et n'exigerai jamais un salaire au-dessus de mon travail.

Admis dans l'intérieur des maisons, mes yeux

ne verront pas ce qui s'y passe, ma langue taira

les secrets qui me seront confiés et mon état ne servira pas

à corrompre les mœurs ni à favoriser le crime.

Respectueux et reconnaissant envers mes Maîtres,

je rendrai à leurs enfants l'instruction

que j'ai reçue de leurs pères.

Que les hommes m'accordent leur estime

si je suis fidèle à mes promesses.

Que je sois couvert d'opprobre

et méprisé de mes confrères

si j'y manque. »

Remerciements

Je tiens à remercier Monsieur RUSCH, Professeur à l'Université de médecine de Tours, qui m'a fait confiance sur le choix du sujet et m'a partagé sa vision transversale. Qu'il soit aussi remercié pour sa bienveillance depuis le début et ses conseils avisés sur cette thèse.

Je remercie Madame GUILLON-GRAMMATICO, Maître de Conférences à l'Université de médecine de Tours pour son encadrement sur ce travail, fruit de plus de trois années enrichissantes. C'est à ses côtés que j'ai pu observer ce que rigueur et précision voulaient signifier.

Pour son accompagnement sur ce travail, je remercie Monsieur DHALLUIN, Assistant Hospitalier Universitaire.

J'adresse tous mes remerciements à Monsieur LINASSIER, Professeur à l'Université de Tours, ainsi qu'à Madame PLICHON, Praticien Hospitalier au Centre Hospitalier Régional Universitaire de Tours, de l'honneur qu'ils m'ont fait en acceptant d'être jurés de cette thèse.

Monsieur PASCO, Docteur en Santé Publique, m'a non seulement initié à l'informatique médicale et aux bases de données lorsque j'étais nouvelle interne, mais aussi prodigué des conseils pour le choix de mon Mastère II dont cette thèse est l'accomplissement. Qu'il en soit remercié.

Je tiens aussi à remercier Monsieur HEITZMANN, Médecin coordonnateur du réseau de cancérologie de la région Centre et Madame LEFEBVRE qui m'ont accueilli pour un stage d'internat au sein de leur équipe d'ONCOCENTRE. C'est grâce à eux que j'ai appris comment concilier recherche et appui aux professionnels, objet de cette thèse.

Merci aussi à Monsieur CUGGIA, Professeur à l'Université de médecine de Rennes, dont les thèmes de recherche ont fortement inspiré ce travail.

Enfin, je tiens à remercier Madame SALPETRIER et Monsieur HERBERT, data managers au centre de données cliniques, qui ont répondu avec pertinence et patience à mes questions régulières sur les entrepôts de données hospitalières.

Un grand merci aussi à tous les membres du service d'information médicale, d'épidémiologie et d'économie de la santé en particulier à Monsieur GABORIT pour ses explications sur les modèles d'apprentissage supervisés.

Dédicaces

A Rémi, mes chers Parents & Grands Parents, Pierre, Jeanne, Hugues, mes précieux amis et co-internes, un grand MERCI du fond du cœur pour tout votre soutien.

Table des matières

Liste des tableaux et figures

Liste des annexes

Liste des abréviations, sigles et acronymes

AJCC	American Joint Committee on Cancer
ALK	Anaplastic Lymphoma Kinase
ANSM	Agence Nationale de Sécurité du Médicament et des produits de santé
API	Application Programmation Interface
ARC	Attaché de recherche clinique
BPC	Bonnes Pratiques Cliniques
CNPC	Carcinome Non à Petites Cellules
CPC	Carcinome à Petites Cellules
EGFR	Epidermal Growth Factor Receptor
eHOP®	Entrepôt des Données Hospitalières
EudraCT	European Union Drug Regulating Authorities Clinical Trials Database
GREPP	Groupement Régional Evaluation des Pratiques Professionnelles
ICH	International Conference on Harmonization
INCa	Institut National contre le Cancer
IQSS	Indicateurs de Qualité et Sécurité des Soins
IRM	Imagerie par Résonnance Magnétique
M	Metastasis
N	Node
NCI	National Cancer Institute
OMS	Organisation Mondiale de la Santé
PreScIOUS	PreScreenning In Oncology Using Data Science
PS OMS	Performance Status OMS
RECF	Registre des Essais Cliniques en Cancérologie en France
RCP	Réunion de Concertation Pluridisciplinaire
RegEx	Regular Expressions
Ri-CDC	Réseau interrégional du Centre de Données Cliniques
SEER	Surveillance, Epidemiology and End Results
T	Tumor
UICC	Union for International Cancer Control

Introduction

Le cancer est la troisième cause de mortalité après les maladies cardio-vasculaires et démences dans les pays occidentaux en 2019 (1). L'évolution des traitements ces vingt dernières années a permis d'améliorer l'espérance de vie des patients atteint de cancer broncho-pulmonaire (2,3). La mortalité rapportée à l'incidence diminue de façon importante (- 6,3 % par an) chez les hommes atteints de carcinomes non à petites cellules (CNPC) (4). La survie spécifique à deux ans de ce type majoritaire a augmenté (+ 40 %) en dix ans et cette évolution, plus importante et deux fois plus rapide que la baisse d'incidence, coïncide avec l'amélioration des techniques diagnostiques (dépistage et diagnostic moléculaire) et thérapeutiques (thérapeutiques ciblées) à partir de 2006 (4,5).

Une progression comparable est attendue dans les prochaines années grâce à l'utilisation de l'immunothérapie. Des survies globales de plus de 36 mois ont été observées grâce aux essais cliniques sur des traitements anticancéreux oraux ciblant des mutations telles les *anti-epidermal growth factor receptor* (EGFR), puis les *anti-anaplastic lymphoma kinase* (ALK) (6). Mieux supportées que les chimiothérapies, ces molécules sont également plus actives à chaque nouvelle génération mise sur le marché (7–9).

Les médecins rapportent être favorables aux essais cliniques et les considèrent comme une source de soins de haute qualité (10). Cependant, moins de 5 % des patients atteints de cancer y participent (11–13) par manque d'essais menés dans leur centre de soin correspondant au type et au stade de leur cancer (14). Le taux de participation peut doubler d'un centre à l'autre, favorisant les grands centres universitaires (14). A côté des facteurs institutionnels prépondérants, la participation aux essais cliniques dépend aussi de facteurs organisationnels tels les ressources en temps-humains et la communication entre centres (15).

En France, l'Institut National contre le Cancer (INCa), groupement d'intérêt public fondé en 2005, est chargé de coordonner la recherche scientifique et la lutte contre le cancer. Il pilote pour le compte des ministères chargés de la santé et de la recherche, le Plan cancer 2014-2019 qui a pour priorités de santé, l'accompagnement des évolutions technologiques et thérapeutiques (Objectif 3) et l'accélération de l'innovation au bénéfice des patients (Objectif 5).

Afin d'informer patients et professionnels de santé des essais cliniques menés en cancérologie, un Registre des Essais Cliniques en cancérologie en France (RECF) a été créé par l'INCa, en collaboration avec l'Agence nationale de sécurité du médicament et des produits de santé (ANSM) définie comme l'autorité compétente pour les recherches impliquant la personne humaine (Article L.1121-1, Code de la santé publique) (16).

Les promoteurs des essais cliniques ont pour obligation de publier les résultats de leurs essais cliniques dans l'European Clinical Trials Database (EudraCT) dans un délai d'un an à compter de la fin de l'essai (17). Le RECF est alimenté par l'EudraCT et le nombre d'essais enregistrés est au moins aussi élevé.

Les protocoles des essais cliniques enregistrés sont consultables sous forme de résumés destinés aux patients et de fiches d'informations scientifiques plus détaillées pour les professionnels (18). Le RECF met à disposition un moteur de recherche multicritères ainsi qu'un module de géolocalisation calculant un itinéraire vers un établissement d'investigation.

Une source importante d'informations concernant la prise en charge diagnostique et thérapeutique des patients atteints de cancer est représentée par les comptes rendus de réunions de concertation pluridisciplinaires (RCP). Les RCP regroupent des professionnels de santé de différentes disciplines dont les compétences sont indispensables pour prendre une décision accordant aux patients la meilleure prise en charge en fonction de l'état de la science. Les modalités d'organisation de la RCP sont définies par l'article D. 6124-131 du Code de la santé publique.

La décision thérapeutique est tracée de façon semi structurée (modules et champs prédéfinis de texte libre) puis est soumise et expliquée au patient (19). Pour aider à la communication entre professionnels, les comptes rendus de RCP sont produits par l'équipe du centre de coordination en cancérologie et gérés dans le dossier communicant de cancérologie du réseau OncoCentre via la base Infocentre. Ils sont également déversés dans l'entrepôts de données hospitalières (eHOP®) du centre de données cliniques (CDC) (20).

Au vu de la grande quantité d'essais cliniques en cours de recrutement et de la grande quantité de patients suivis dans chaque centre hospitalier et de lutte contre le cancer, la numérisation et la structuration de ces informations est une opportunité pour développer un outil automatisé pouvant aider au recrutement. Comparée à la méthode manuelle, détecter automatiquement les sujets éligibles permet de réduire la charge de travail des équipes de recherche tout en augmentant l'efficacité de la tâche de pré-sélection (21–25).

Les critères d'éligibilité des sujets peut être réalisée par des requêtes sur une chaîne de caractères (lettres, chiffres, symboles, ponctuations) selon une syntaxe spécifique aux regEx à implémenter (26). Issues des théories mathématiques des langages de programmation informatique des années 1940, les expressions régulières ou règles d'experts (regEx) ou encore motifs, sont une technique de Traitement Automatique du Langage (TAL) (27).

Cette étude a pour objectif de proposer un outil d'aide à la pré-sélection de sujets à l'entrée dans les essais cliniques répondant à la question : Quel(s) sujet(s) pour quel(s) protocole(s) ? Ce travail consiste à implémenter puis évaluer un outil basé sur des regEx permettant d'apparier les comptes rendus de RCP de patients atteints de cancer du poumon avec les fiches des protocoles d'essais cliniques enregistrées dans le RECF.

Méthodologie

I. Détection des caractéristiques des sujets à partir des RCP

Les comptes rendus de RCP d'oncologie thoracique réalisés entre le 1er janvier et le 15 juillet 2021, stockés au sein de l'entrepôt de notre établissement et déidentifiés ont été étudiés. Ils étaient peu structurés car rédigés en texte libre (28,29). Des variations importantes (syntaxe, abréviation, orthographe) ont été relevées et les textes ont dû être normalisés, prérequis indispensable pour l'utilisation de TAL.

L'algorithme PreScIOUS[1] a été utilisé pour rechercher au sein des comptes rendus de RCP huit caractéristiques concernant des critères d'éligibilités liées au patient et à son cancer. Ces critères d'éligibilité concernaient l'âge au moment de la RCP, le score du statut de performance de l'OMS (PS OMS), la ligne de traitement, le stade TNM (critères AJCC/UICC) imputé, le type de carcinome à petites cellules (CPC) ou non, le type histologique précis et la présence ou non de mutations des gènes EGFR et ALK. Concernant le type histologique, les tumeurs les plus rares (sarcome, tumeurs myoépithéliales…) étaient regroupées dans la catégorie « autres ».

Un même motif isolé pouvant correspondre à différentes significations (ex : « t1 » pour « stade tumoral 1 » mais aussi « vertèbre thoracique numéro 1 » ou « IRM pondérée en t1 »), des regEx précises étaient implémentées pour exclure certains motifs (ex : « t4 » est considéré comme stade tumoral 4 sauf s'il est précédé du mot « vertèbre »).

Le stade TNM a été inféré à partir des informations sur les T, N et M extraites des RCP (30). Si l'information recherchée n'était pas suffisamment précisée dans le compte rendu de RCP, le stade le moins avancé était retenu par défaut (ex : « M1 » était classé comme « M1a »).

II. Détection des critères d'éligibilité à partir des protocoles

Une requête via la page du RECF sur le site de l'INCa a permis de sélectionner les protocoles selon cinq critères : essais de phase III, en cours d'inclusion (ou inclusions à venir), menés dans le Grand Ouest (Centre-Val-de-Loire, Bretagne et Pays-de-Loire), concernant les cancers du poumon à petites cellules ou non à petites cellules, chez l'adulte.

[1] Ansoborlo M, Dhalluin T, Gaborit C, Cuggia M, Grammatico-Guilllon L. Prescreening in Oncology Using Data Sciences: The PreScIOUS Study. Stud Health Technol Inform. 27 mai 2021;281:123-7.

L'outil était basé sur des regEx appliquées sur des extractions automatiques de données depuis la page internet publique de l'INCa (31,32). Au sein des fiches détaillées des protocoles sélectionnés, en plus des huit caractéristiques du profil patient-cancer, étaient extraites les coordonnées du coordonnateur référent.

III. Analyses statistiques

Les analyses ont été réalisées via le logiciel R version 3.6.0 (33).

a. Calcul des paramètres de performance de l'outil

Pour chaque information extraite depuis les protocoles, étaient comparés les résultats des regEx (« *positifs* » ou « *négatifs* ») à ceux de référence (« *vrai* » ou « *faux* »). Les taux de concordance (vrais positifs, faux positifs, vrais négatifs et faux négatifs) ainsi que les paramètres de performance en termes de taux de précision, de rappel ont été calculés :

$$\boldsymbol{Précision} = \frac{Vrais\ positifs}{Faux\ positifs + Vrais\ positifs}$$

$$\boldsymbol{Rappel} = \frac{Vrais\ positifs}{Faux\ négatifs + Vrais\ positifs}$$

Ces deux paramètres sont complémentaires car la précision correspond à la confiance que l'on peut avoir en l'outil alors que le rappel correspond à sa puissance de détection. Dans le cadre de la pré-sélection, chercher à optimiser le rappel permet de minimiser le nombre d'exclusions de sujets potentiellement éligibles à tort (faux négatifs).

Pour obtenir un indicateur de performance global unique pour chaque critère à extraire, la F-mesure a été utilisée. Elle correspond à la moyenne harmonique entre la précision et le rappel :

$$\boldsymbol{F-mesure} = \frac{2 * Précision * Rappel}{Précision + Rappel}$$

b. Statistiques descriptives

Une étude descriptive de la répartition des modalités des huit critères au sein de l'échantillon des protocoles et celui des comptes rendus de RCP a été réalisée en monocentrique. Ont été calculés la moyenne et l'écart type du critère quantitatif ainsi que les effectifs et proportions pour les critères qualitatifs.

Pour le critère quantitatif « âge » et les critères catégoriels ordinaux « stade TNM » et « score OMS », des seuils minimaux et maximaux étaient systématiquement spécifiés au sein des protocoles. Ces critères d'éligibilité correspondaient à des seuils à ne pas dépasser et non une valeur unique à vérifier. Ainsi pour chacun de ces trois critères, deux valeurs étaient extraites des protocoles alors qu'une valeur unique était extraite des comptes rendus de RCP (ex : âge exact du patient au moment de la RCP).

Le pouvoir discriminant de chaque critère d'éligibilité a été estimé par le nombre de sujets supplémentaires potentiellement inclus lorsque celui-ci n'était pas vérifié, appelé « gain d'inclusion par critère ». Cet indicateur était la différence entre le nombre moyen de sujets éligibles (en excluant ce critère) après vérification des sept autres critères et le nombre moyen de sujets éligibles après vérification de l'ensemble des huit critères.

c. Critère de jugement principal

Le critère de jugement principal était l'estimation des paires sujets-protocoles à partir des comptes rendus de RCP de l'entrepôt de données hospitalières et des essais cliniques publiés sur le RECF. Les informations extraites de ces deux sources ont été croisées puis la proportion de comptes rendus de RCP du CHRU de Tours appariés et celle des protocoles ouverts aux inclusions en 2021 proposables ont été estimées.

Un appariement était compté lorsque l'ensemble des informations présentes au sein du compte rendu de RCP correspondaient aux critères d'éligibilité spécifiés dans le protocole. Par défaut, lorsque l'information n'était pas renseignée au sein du compte rendu de RCP, le sujet était considéré comme apparié car son dossier pouvait être évalué dans le cadre d'une pré-sélection.

Résultats

Seize protocoles ont été extraits du registre de l'INCa, concernant les essais cliniques de phase III étudiant les traitements contre les cancers trachéo-bronchiques chez l'adulte ouverts à l'inclusion en juillet 2021 menés en régions Bretagne (n=11), Centre-Val de Loire (n=5) et Pays de Loire (n=12) (figure 1). Les informations concernant les sites investigateurs sont renseignées dans la majorité des fiches d'information des protocoles publiées via le RECF.

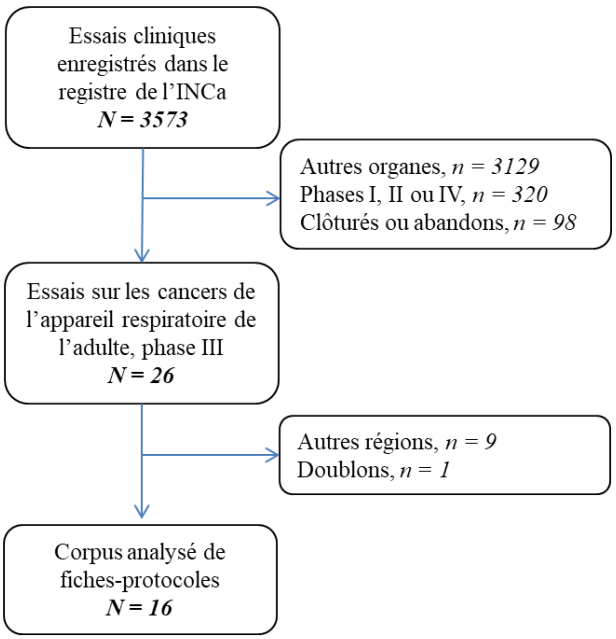


Figure 1. Diagramme de flux des protocoles

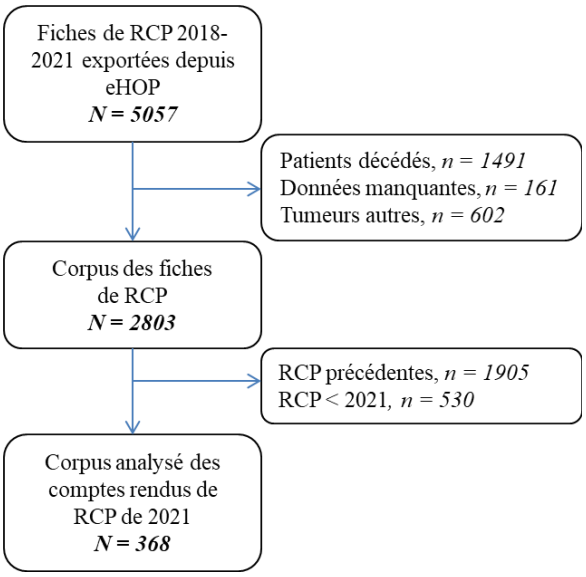


Figure 2. Diagramme de flux des comptes rendus de RCP

Un échantillon de 368 comptes rendus de RCP dont l'enregistrement a été effectué entre le 7 janvier et le 15 juillet 2021 a été constitué à partir d'une extraction de l'entrepôt de données hospitalières (figure 2). Lorsque plusieurs comptes rendus concernaient un même patient, seul le plus récent était inclus dans l'analyse pour constituer un échantillon de 368 comptes rendus - sujets potentiels à pré-sélectionner. Les comptes rendus de RCP concernant des tumeurs autres que les tumeurs trachéo-bronchiques ou ayant l'un des champs « histoire de la maladie » ou « antécédents » vide étaient exclues de l'analyse.

I. Paramètres de performance de l'outil

La performance de l'outil d'extraction des critères d'éligibilité s'élevait à 86 % de F-mesure en moyenne. Ce taux était variable selon les critères extraits des protocoles du RECF (tableau 1). L'outil atteignait des taux de précision et de rappel les plus élevés pour les mutations EGFR et ALK, l'âge et l'histologie. Les performances les plus faibles étaient observées pour le statut de performance de l'OMS et le stade TNM.

Tableau 1. Paramètres de performance de l'outil d'extraction d'informations des protocoles

	Rappel %	**Précision %**	**F-mesure %**
Age			
Minimum à l'inclusion	83	96	88
Maximum à l'inclusion	100	100	100
PS OMS			
Minimum à l'inclusion	50	47	48
Maximum à l'inclusion	100	100	100
Traitement de 1ère ligne	89	64	67
Stade TNM			
Minimum à l'inclusion	72	62	66
Maximum à l'inclusion	100	100	100
Histologie			
Type	100	100	100
Sous-type	82	78	73
Mutation			
EGFR	100	100	100
ALK	100	100	100
Performance moyenne	89	86	86

II. Description des caractéristiques extraites par l'outil

Les critères les moins spécifiés par les protocoles étaient le seuil minimal éligible pour le PS OMS (6 %), la présence de mutations EGFR (6 %) et ALK (13 %) ainsi que le traitement de 1ère ligne (19 %). L'âge maximal éligible, le seuil maximal du PS de l'OMS et le type histologique (à petites cellules ou non) étaient systématiquement renseignés.

Les variables les plus rarement renseignées au sein des comptes rendus de RCP étaient les mutations ALK (15 %) et EGFR (38 %) (tableau 2). Au sein de notre échantillon de comptes rendus de RCP, les deux grands types histologiques étaient équitablement représentés (53 % de CNPC et 47 % de CPC), contrairement à la répartition épidémiologique en population générale (29).

Tableau 2. Description des comptes rendus de RCP et des protocoles du RECF

Caractéristiques extraites		RCP-sujets (N= 368)	Protocoles (N=16)		
			Seuil minimal	*Valeur exacte*	*Seuil maximal*
Région, n (%)	Bretagne	0 (0,0)		3 (18,8)	
	Centre-Val de Loire	368 (100)		5 (31,2)	
	Pays de la Loire	0 (0,0)		8 (50,0)	
Sex-ratio		2,0		*NR*	
Traitement de 1ère ligne, n (%)	Oui	99 (26,9)		2 (12,5)	
	Non	269 (73,1)		1 (6,2)	
	NR	0 (0,0)		13 (81,2)	
Carcinome, n (%)	A petites cellules	173 (47,0)		0 (0,0)	
	Non à petites cellules	195 (53,0)		16 (100)	
Type histologique, n (%)	Adénocarcinome	176 (47,8)		1 (6,2)	
	Epidermoïde	76 (20,7)		5 (31,2)	
	Neuroendocrine	41 (11,1)		0 (0,0)	
	Autre	25 (6,8)		0 (0,0)	
	NR	50 (13,6)		10 (62,5)	
Gène EGFR, n (%)	Muté	23 (6,2)		1 (6,2)	
	Non muté	117 (318)		0 (0,0)	
	Non testé	2 (0,5)		0 (0,0)	
	NR	226 (61,4)		15 (93,8)	
Gène ALK, n (%)	Muté	8 (2,2)		2 (12,5)	
	Non muté	46 (12,5)		0 (0,0)	
	Non testé	1 (0,3)		0 (0,0)	
	NR	313 (85,0)		14 (87,5)	
Age, moyenne (ET)		66,9 (10,1)	27,7 (20,9)		79,5 (13,4)
Score PS de l'OMS, n (%)	0	181 (49,2)	0 (0,0)		0 (0,0)
	1	139 (37,8)	0 (0,0)		11(68,8)
	2	32 (8,7)	1 (6,2)		4 (25,0)
	3	9 (2,4)	0 (0,0)		1 (6,2)
	4	7 (1,9)	0 (0,0)		0 (0,0)
	NR	0 (0,0)	15 (93,8)		0 (0,0)
Stade TNM, n (%)	Ia	2 (0,5)	1 (6,2)		0 (0,0)
	IIa	10 (2,7)	3 (18,8)		1 (6,2)
	IIIa	17 (4,6)	1 (6,2)		0 (0,0)
	IIIb	18 (4,9)	3 (18,8)		3 (18,8)
	IIIc	1 (0,3)	0 (0,0)		1 (6,2)
	IVa	76 (20,7)	4 (25,0)		7 (43,8)
	IVb	55 (14,9)	0 (0,0)		0 (0,0)
	NR	189 (51,4)	4 (25,0)		4 (25,0)

NR : non renseigné, ET : écart-type.

III. Description des paires comptes rendus de RCP et protocoles

Parmi les 368 sujets du corpus de comptes rendus de RCP, 109 (29,6 %) étaient potentiellement éligibles sur l'ensemble des 16 études ouvertes à l'inclusion en juillet 2021 dans le Grand Ouest (tableau 3). Au maximum, 9 protocoles pouvaient être proposés à un potentiel sujet (annexes, figure A.2).

Tableau 3. Distribution des sujets selon leur nombre maximal de protocoles potentiels

		Nombre maximal de protocoles appariés				
		Aucun	***Un à trois***	***Quatre à six***	***Sept et plus***	***Total***
Sujets,	n	259	30	54	25	368
	(%)	(70.4)	(8.1)	(14.7)	(6.8)	(100)

Au maximum, un protocole pouvait inclure jusqu'à 65 patients (figure 3). La majorité des protocoles (n=9) pouvaient inclure au moins 28 patients. Seulement trois protocoles ne correspondaient à aucun sujet (annexes, tableau A.1). En moyenne sur les seize protocoles, 7,3 % des patients évalués étaient potentiellement éligibles (annexes, figure A.1). Cette proportion augmentait peu (+ 2,2 points) si l'appariement concernait également les patients dont la dernière RCP datait de 3 ans au plus (annexes, figure A.1).

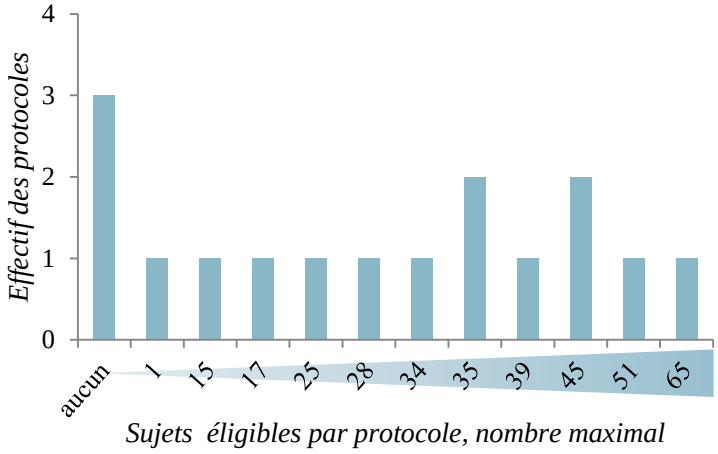


Figure 3. Répartition des protocoles des essais ouverts en juillet 2021 dans le Grand Ouest, selon leur potentiel d'inclusion à partir du CHRU de Tours

Parmi l'ensemble des critères à vérifier, le stade TNM permettait de discriminer le plus efficacement les patients non éligibles devant le type histologique et le statut de performance de l'OMS (tableau A.2).

Discussion

Un outil d'aide à la pré-sélection des patients à l'entrée dans les essais cliniques a été implémenté puis évalué en utilisant un algorithme de TAL. Avec une performance s'élevant à 88 % de F-mesure en moyenne, l'outil peut apparier des comptes rendus de RCP et des protocoles d'essais thérapeutiques contre le cancer du poumon (34). La sensibilité de l'outil pour extraire les caractéristiques des patients était plus élevée que celles retrouvées dans la littérature (35,36).

En 2021, 7,3 % des patients étaient potentiellement éligibles en moyenne sur les 16 études ouvertes aux inclusions dans le Grand Ouest. Ce résultat concorde avec ceux rapportés par les oncologues estimant inclure jusqu'à 7 % de leurs patients dans les essais cliniques (15). En moyenne, une étude pouvait être proposée par patient avec toutefois, de fortes disparités selon les patients. Pour la majorité d'entre eux, aucun protocole ne pouvait être proposé. Aucun protocole ne concernait les CPC, possiblement en lien avec les résultats décourageants des essais cliniques sur ce type histologique (37).

Le type histologique du cancer était le critère le plus discriminant devant la présence de mutation ALK, le PS de l'OMS et le stade du cancer. La littérature rapporte que l'absence de proposition aux patients de participer à un essai clinique était liée au PS OMS faible et au type histologique à petites cellules (38).

Dans la majorité des cas, les performances de l'outil étaient parallèles aux taux de renseignement des informations. La qualité des renseignements au sein des comptes rendus de RCP est optimisée grâce à leur standardisation et validation par les équipes du centre de coordination en cancérologie. L'aide à l'enregistrement et la validation par le centre de coordination en cancérologie d'une part et les audits internes et externes (ex : IQSS RCP) ou les évaluations proposées par le groupe régional d'évaluation des pratiques professionnelles (GREPP) d'autre part, permettent d'assurer la qualité des informations renseignées au sein des comptes rendus de RCP (39).

Pour estimer le nombre potentiel de sujets pouvant bénéficier d'une pré-sélection, l'évaluation de l'outil était basée sur les données de 2021, limitant l'échantillon à 16 protocoles ouverts actuellement ou prochainement dans le Grand Ouest. Au vu de la structuration des données du RECF, un travail sur un plus grand échantillon de critères d'inclusion serait possible pour améliorer la robustesse des algorithmes.

Après un prétraitement du texte libre, des modèles d'apprentissage automatique (*machine learning*) pourraient être implémentés pour augmenter les performances de l'extraction d'information concernant les stades des cancers du poumon notamment (40). Les informations difficilement extraites (ex : résultats des recherches de mutations) pourraient être reconnues par des systèmes mixtes alliant regEx et apprentissage automatique (34).

L'utilisation de regEx permet une meilleure interprétation et transparence des algorithmes. Ceux-ci seront appelés à évoluer au fur et à mesure des avancées de la recherche, aussi, la simplicité de l'implémentation de nouveaux critères standardisés à extraire (ex : nouvelles mutations, résultats de biologie…) sera nécessaire pour s'adapter aux progrès thérapeutiques.

Actuellement, le contrôle des critères d'inclusion est assuré par les attachés de recherche clinique (ARC) mais le manque de temps et de ressources qui leurs sont dédiés représente la barrière principale au recrutement des sujets (41). En pratique, cet outil d'aide à la pré-sélection considère le sujet comme éligible lorsque l'information n'est pas renseignée au sein de la fiche de RCP, la sélection finale étant faite par l'équipe médicale et de recherche. De plus, l'état clinique du patient au moment de la proposition d'inclusion peut avoir évolué par rapport à celui renseigné lors de la dernière RCP.

L'outil s'applique volontairement aux cas de cancers broncho-pulmonaires uniquement. Le choix des cancers broncho-pulmonaires tient d'une part à leur fréquence et au fort potentiel d'inclusion dans des protocoles de recherche et d'autre part, à la qualité des RCP en pneumo-oncologie (42).

L'évaluation de l'outil est conditionnée par la fiabilité de l'annotation manuelle initiale des deux échantillons (protocoles et comptes rendus de RCP). Ces étiquetages sont considérés comme étant les données de référence et sont utilisés pour estimer les performances de l'outil. Les annotations manuelles des échantillons pour les différents critères pourraient être croisées avec celles d'un spécialiste afin de tester la concordance inter-annotateurs.

Par ailleurs, la liste des critères d'inclusion automatisables peut encore être alimentée avec le statut vital. Elargir la fouille d'information concernant les mutations ALK et EGFR aux comptes rendus d'anatomo-pathologie représente un fort potentiel d'amélioration de l'exhaustivité. La connaissance des antécédents du sujet, de son état général dans les observations médicales pourrait affiner la pré-sélection des sujets (43) de même que la vérification de l'absence de comorbidité (44,45). Les données de biologie concernant les fonctions hématologique, rénale ou hépatique pourraient également être discriminantes pour l'entrée dans les essais cliniques.

Appliqué en vie réelle, un logiciel d'aide au recrutement pourrait permettre une augmentation des taux d'inclusion aux essais thérapeutiques (23). Par un système de notification automatique, l'équipe de recherche pourrait demander à recevoir la liste des essais du registre pouvant potentiellement inclure les sujets dont les dossiers ont été étudiés en RCP. Pour éviter une « sur-notification » d'essais positifs à tort, un arbitrage sera à trouver entre des critères d'appariement inclusifs (positif par défaut si valeur manquante) ou restrictifs (négatif par défaut si valeur manquante).

En perspective, le renseignement des sujets pressentis à l'inclusion dans les essais menés en région au sein de la grille d'activité DCC d'OncoCentre par les techniciens d'étude cliniques (TEC) et ARC pourrait en être facilité (46). Cependant, le temps de traitement de l'outil étant de plusieurs minutes (plus important pour traiter le registre que les comptes rendus de RCP), une

validation en routine serait nécessaire auprès des TEC et ARC afin de comparer le gain en terme d'allégement de la charge de travail avec ceux estimés pour des outils similaires (35).

Par ailleurs, la mise en production d'une interface de programmation d'application (*API*) entre le RECF et l'entrepôt eHOP® permettrait un usage en temps réel de l'outil. Cependant en France, la commission nationale de l'informatique et des libertés (CNIL) identifie les coordonnées des investigateurs collectées comme des données sensibles. Les investigateurs ont donné leur consentement pour le traitement de leurs données par l'INCa spécifiquement. Recueillir leur accord serait nécessaire pour réutiliser leurs données dans le cadre de cet outil en vie réelle.

Développer PreScIOUS au sein d'eHOP® permettrait son déploiement à l'échelle du Ouest Data Hub afin d'aider les équipes de recherche des différents centres du Grand Ouest à inclure de nouveaux sujets dans les essais cliniques en pneumo-oncologie. Les comptes rendus médicaux étant soumis à une variabilité (rédaction, format) entre les différents centres de soin, la validité externe de l'outil doit être estimée. Tester l'outil sur les fiches RCP d'autres centres sera nécessaire pour l'interopérabilité des entrepôts de données concernant les recrutements multicentriques (47).

Ce travail a permis de démontrer la faisabilité de construire et d'évaluer un outil automatisé d'aide à la pré-sélection des sujets potentiellement éligibles à l'entrée dans un essai thérapeutique en oncologie, sur le modèle d'un cancer prévalent et toujours à fort enjeu de santé publique.

Afin de proposer in fine, l'entrée dans un essai thérapeutique aux patients suivis pour un cancer du poumon, automatiser la fouille de leurs dossiers médicaux diminuerait la charge de travail avec une performance au moins aussi élevée que les procédures actuelles, en gagnant du temps Homme. Des travaux supplémentaires sont encore nécessaires pour la validation complète de ce type d'outil, mais l'aide apportée par les outils innovants du Big Data en santé est déjà en ligne de mire…

Bibliographie

1. Les 10 principales causes de mortalité [Internet]. [cité 19 août 2021]. Disponible sur: https://www.who.int/fr/news-room/fact-sheets/detail/the-top-10-causes-of-death

2. Henley SJ, Thomas CC, Lewis DR, Ward EM, Islami F, Wu M, et al. Annual report to the nation on the status of cancer, part II: Progress toward Healthy People 2020 objectives for 4 common cancers. Cancer. 15 mai 2020;126(10):2250-66.

3. Francisci S, Minicozzi P, Pierannunzio D, Ardanaz E, Eberle A, Grimsrud TK, et al. Survival patterns in lung and pleural cancer in Europe 1999-2007: Results from the EUROCARE-5 study. Eur J Cancer. oct 2015;51(15):2242-53.

4. Howlader N, Forjaz G, Mooradian MJ, Meza R, Kong CY, Cronin KA, et al. The Effect of Advances in Lung-Cancer Treatment on Population Mortality. N Engl J Med. 13 août 2020;383(7):640-9.

5. Street W. Cancer Facts & Figures 2020. 1930;76.

6. Peters S, Camidge DR, Shaw AT, Gadgeel S, Ahn JS, Kim D-W, et al. Alectinib versus Crizotinib in Untreated ALK-Positive Non-Small-Cell Lung Cancer. N Engl J Med. 31 août 2017;377(9):829-38.

7. Soria J-C, Ohe Y, Vansteenkiste J, Reungwetwattana T, Chewaskulyong B, Lee KH, et al. Osimertinib in Untreated EGFR-Mutated Advanced Non-Small-Cell Lung Cancer. N Engl J Med. 11 janv 2018;378(2):113-25.

8. Soria J-C, Tan DSW, Chiari R, Wu Y-L, Paz-Ares L, Wolf J, et al. First-line ceritinib versus platinum-based chemotherapy in advanced ALK-rearranged non-small-cell lung cancer (ASCEND-4): a randomised, open-label, phase 3 study. Lancet. 4 mars 2017;389(10072):917-29.

9. Soria J-C, Felip E, Cobo M, Lu S, Syrigos K, Lee KH, et al. Afatinib versus erlotinib as second-line treatment of patients with advanced squamous cell carcinoma of the lung (LUX-Lung 8): an open-label randomised controlled phase 3 trial. Lancet Oncol. août 2015;16(8):897-907.

10. T A-H, Ea C, Tr H, C S, Jf P, D J, et al. The Effect of Receiving Treatment Within a Clinical Trial Setting on Survival and Quality of Care Perception in Advanced Stage Non-Small Cell Lung Cancer. Am J Clin Oncol. 1 avr 2016;39(2):126-31.

11. An American Society of Clinical Oncology and Institute of Medicine Workshop, Institute of Medicine, National Cancer Policy Forum, Board on Health Care Services. Implementing a National Cancer Clinical Trials System for the 21st Century: Second Workshop Summary [Internet]. Washington (DC): National Academies Press (US); 2013 [cité 6 août 2021]. Disponible sur: http://www.ncbi.nlm.nih.gov/books/NBK202108/

12. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. JAMA. 9 juin 2004;291(22):2720-6.

13. Tejeda HA, Green SB, Trimble EL, Ford L, High JL, Ungerleider RS, et al. Representation of African-Americans, Hispanics, and whites in National Cancer Institute cancer treatment trials. J Natl Cancer Inst. 19 juin 1996;88(12):812-6.

14. Unger JM, Cook E, Tai E, Bleyer A. The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. Am Soc Clin Oncol Educ Book. 2016;35:185-98.

15. Organizational Barriers to Physician Participation in Cancer Clinical Trials [Internet]. AJMC. [cité 6 août 2021]. Disponible sur: https://www.ajmc.com/view/jul05-2081p413-421

16. Titre II : Recherches impliquant la personne humaine (Articles R1121-1 à R1125-26) - Légifrance [Internet]. [cité 6 août 2021]. Disponible sur: https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006072665/LEGISCTA000006160948/#LEGISCTA000034773672

17. Clinical Trials Register [Internet]. [cité 6 août 2021]. Disponible sur: https://www.clinicaltrialsregister.eu/ctr-search/search

18. A propos du Registre - Le registre des essais cliniques [Internet]. [cité 6 août 2021]. Disponible sur: https://www.e-cancer.fr/Professionnels-de-sante/Le-registre-des-essais-cliniques/A-propos-du-Registre

19. GuideBonUsageDCC.pdf [Internet]. [cité 6 août 2021]. Disponible sur: https://oncocentre.org/wp-content/uploads/GuideBonUsageDCC.pdf

20. OncoCentre – Réseau de Cancérologie de la Région Centre » Présentation de l'outil [Internet]. [cité 6 août 2021]. Disponible sur: https://oncocentre.org/outils/dcc/presentation/

21. Grundmeier RW, Swietlik M, Bell LM. Research subject enrollment by primary care pediatricians using an electronic health record. AMIA Annu Symp Proc. 11 oct 2007;289-93.

22. Nkoy FL, Wolfe D, Hales JW, Lattin G, Rackham M, Maloney CG. Enhancing an existing clinical information system to improve study recruitment and census gathering efficiency. AMIA Annu Symp Proc. 14 nov 2009;2009:476-80.

23. Treweek S, Pearson E, Smith N, Neville R, Sargeant P, Boswell B, et al. Desktop software to identify patients eligible for recruitment into a clinical trial: using SARMA to recruit to the ROAD feasibility trial. Inform Prim Care. 1 janv 2010;18(1):51-8.

24. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform. juin 2010;43(3):451-67.

25. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc. janv 2015;22(1):166-78.

26. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc. 2010;17(4):440-5.

27. Kleene SC. Representation of Events in Nerve Nets and Finite Automata [Internet]. Automata Studies. (AM-34), Volume 34. Princeton University Press; 2016 [cité 6 août 2021]. Disponible sur: https://www.degruyter.com/document/doi/10.1515/9781400882618-002/html

28. Delamarre D, Bouzille G, Dalleau K, Courtel D, Cuggia M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. Stud Health Technol Inform. 2015;210:702-6.

29. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics Initiatives: Two National Projects to Promote Data Sharing in Healthcare. Yearb Med Inform. août 2019;28(1):195-202.

30. Detterbeck FC, Boffa DJ, Kim AW, Tanoue LT. The Eighth Edition Lung Cancer Stage Classification. Chest. janv 2017;151(1):193-203.

31. DeVito NJ, Richards GC, Inglesby P. How we learnt to stop worrying and love web scraping. Nature. 8 sept 2020;585(7826):621-2.

32. RSelenium package - RDocumentation [Internet]. [cité 19 août 2021]. Disponible sur: https://www.rdocumentation.org/packages/RSelenium/versions/1.7.7

33. R: The R Project for Statistical Computing [Internet]. [cité 27 août 2021]. Disponible sur: https://www.r-project.org/

34. Ansoborlo M, Dhalluin T, Gaborit C, Cuggia M, Grammatico-Guilllon L. Prescreening in Oncology Using Data Sciences: The PreScIOUS Study. Stud Health Technol Inform. 27 mai 2021;281:123-7.

35. Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, Kaiser M, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility Pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak. déc 2015;15(1):28.

36. McCowan I, Moore D, Fry M-J. Classification of Cancer Stage from Free-text Histology Reports. :4.

37. Mamdani H, Induru R, Jalal SI. Novel therapies in small cell lung cancer. Transl Lung Cancer Res. oct 2015;4(5):533-44.

38. Gonzalez LE, Sutton SK, Pratt C, Gilbertson M, Antonia S, Quinn GP. The Bottleneck Effect in Lung Cancer Clinical Trials. J Cancer Educ. sept 2013;28(3):488-93.

39. PosterOnco-IQSS_RCP-A4-vf.pdf [Internet]. [cité 27 août 2021]. Disponible sur: https://oncocentre.org/wp-content/uploads/PosterOnco-IQSS_RCP-A4-vf.pdf

40. McCowan IA, Moore DC, Nguyen AN, Bowman RV, Clarke BE, Duhig EE, et al. Collection of Cancer Stage Data by Classifying Free-text Medical Reports. J Am Med Inform Assoc. 2007;14(6):736-45.

41. Isaksson E, Wester P, Laska AC, Näsman P, Lundström E. Identifying important barriers to recruitment of patients in randomised clinical studies using a questionnaire for study personnel. Trials. 30 oct 2019;20(1):618.

42. Panorama des cancers en France_2021.pdf [Internet]. [cité 7 août 2021]. Disponible sur: https://www.e-cancer.fr/pdf_inca/preview/303372/4327939/file/Panorama%20des%20cancers%20en%20France_2021.pdf

43. Association of Patient Comorbid Conditions With Cancer Clinical Trial Participation | Oncology | JAMA Oncology | JAMA Network [Internet]. [cité 6 août 2021]. Disponible sur: https://jamanetwork.com/journals/jamaoncology/article-abstract/2720475

44. Fouad MN, Lee JY, Catalano PJ, Vogt TM, Zafar SY, West DW, et al. Enrollment of Patients With Lung and Colorectal Cancers Onto Clinical Trials. JOP. 1 mars 2013;9(2):e40-7.

45. Barriers to recruiting underrepresented populations to cancer clinical trials: A systematic review - Ford - 2008 - Cancer - Wiley Online Library [Internet]. [cité 19 août 2021]. Disponible sur: https://acsjournals.onlinelibrary.wiley.com/doi/10.1002/cncr.23157

46. 1.1-Les-outils-Poster_CNRC2016-DéploiementRechercheClinique-DCC.pdf [Internet]. [cité 19 août 2021]. Disponible sur: https://oncocentre.org/wp-content/uploads/1.1-Les-outils-Poster_CNRC2016-D%C3%A9ploiementRechercheClinique-DCC.pdf

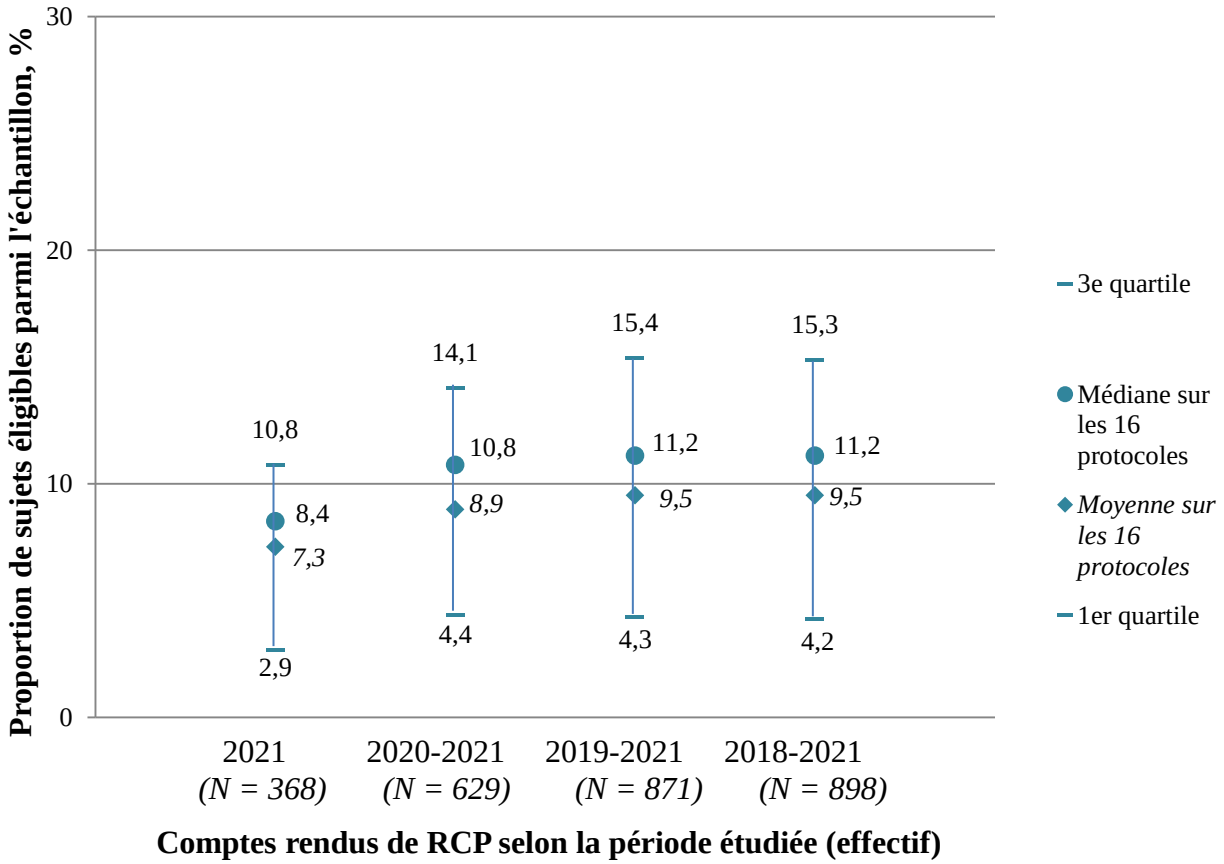
47. Goulart BHL, Silgard ET, Baik CS, Bansal A, Sun Q, Durbin EB, et al. Validity of Natural Language Processing for Ascertainment of EGFR and ALK Test Results in SEER Cases of Stage IV Non-Small-Cell Lung Cancer. JCO Clin Cancer Inform. mai 2019;3:1-15.

Annexes

Tableau A.1. Distribution des protocoles selon leur nombre maximal de sujets éligibles

		Nombre maximal de sujets éligibles					
		Aucun	***[1 ; 15]***	***[16 ; 30]***	***[31 ; 45]***	***Plus de 50***	***Total***
Protocoles,	n	3	2	3	6	2	16
	(%)	(18,7)	(12,5)	(18,7)	(37,5)	(12,5)	(100)

Figure A.1. Appariements entre les protocoles publiés sur le RECF ouverts au recrutement en 2021 dans le Grand Ouest et les sujets appariés, selon la date de RCP.



<u>Interprétation</u> : Le 1er quartile est la valeur de la série qui sépare les 25 % inférieurs des données, le 3e quartile est celle qui sépare les 75 % inférieurs. Exemple : au sein de l'échantillon des 16 protocoles, un quart d'entre eux pouvait être apparié au minimum à 11 % des 368 comptes rendus de RCP réalisés en 2021.

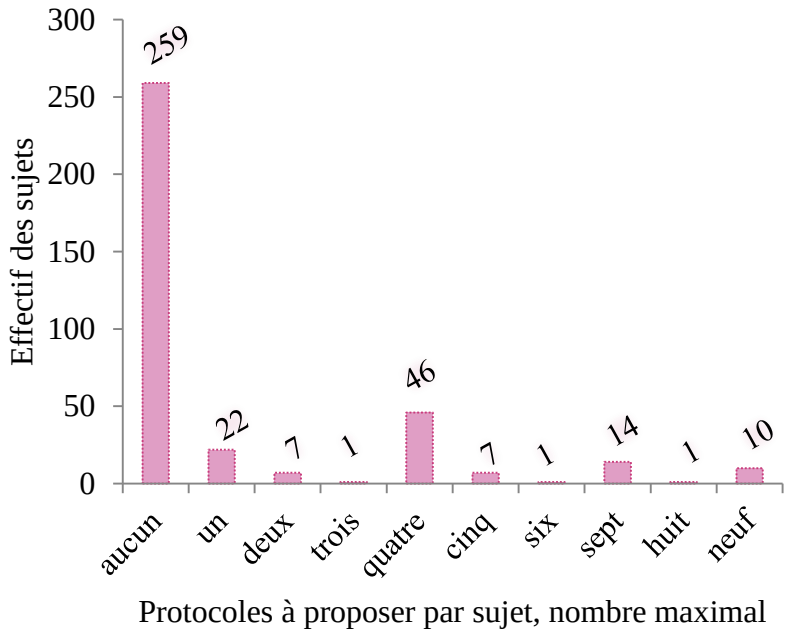


Figure A.2. Répartition des patients selon leur potentiel d'éligibilité parmi les protocoles ouverts en 2021 dans le Grand Ouest

Tableau A.2. Classement des critères d'éligibilité des protocoles selon leur pouvoir discriminant au sein des RCP réalisées en 2021

Critères vérifiés pour l'appariement		Sujets par protocole		
		Nombre moyen	**Part des sujets**	**Gain**
		n	*(%)*	*n*
Tous		27	(7,3)	
Tous sauf	*1ère ligne de traitement*	28	(7,6)	+ 1
	Age	29	(7.9)	+ 2
	Mutation EGFR	29	(7.9)	+ 2
	PS OMS	30	(8.1)	+ 3
	Stade TNM	31	(8.4)	+ 4
	Mutation ALK	33	(8.9)	+ 6
	Histologie	124	(33.7)	+ 97
Aucun		368	(100)	

<u>Interprétation</u> : Le gain par critère est la différence entre le nombre de sujets éligibles en vérifiant l'ensemble des critères, et celui en vérifiant tous les critères sauf celui-ci. Le critère « histologie » concerne le sous-type histologique (ex : épidermoïde, adénocarcinome…), l'ensemble des protocoles concernant des CNPC.

Vu, le Directeur de Thèse,

Tours, le

Vu, le Doyen de la Faculté de Médecine de Tours,

Tours, le

université de TOURS | Faculté de médecine

Thèse 2020/2021

D O C T O R A T e n M É D E C I N E

Diplôme d'État

D. É.S. de « Santé Publique »
Présentée et Soutenue le 20 Octobre 2021.
Dépôt de sujet de thèse, proposition de jury,

NOM : ANSOBORLO
Prénoms : Marie, Françoise, Louise
Date de naissance : 2 Avril 1992 **Nationalité :** française
Lieu de naissance : Bordeaux (33)
Domicile : 60, Rue de la Californie, 37000 Tours **Téléphone :** 06/27/59/14/45

Directeur de Thèse : Docteur Leslie GUILLON-GRAMMATICO

Titre de la Thèse : **Outil d'aide à la pré-sélection dans les essais cliniques en pneumo-oncologie : apparier les fiches RCP et les protocoles du registre français**

JURY

Président : Professeur Emmanuel RUSCH, Epidémiologie, économie de la santé et prévention, Faculté de Médecine – Tours
Membres :
Professeur Claude LINASSIER, Cancérologie, radiothérapie, Faculté de Médecine – Tours
Docteur Leslie GUILLON-GRAMMATICO, Epidémiologie, économie de la santé et prévention, MCU-PH, HDR, Faculté de Médecine – Tours
Docteur Chloé PLICHON, Biopharmacie Clinique Oncologique, PH, CHU – Tours

Avis du Directeur de Thèse
à Tours, le 20.09.2021

Avis du Directeur de l'U.F.R.
à Tours, le 1.10.21

ANSOBORLO Marie

31 pages – 3 tableaux – 3 figures – 4 annexes.

Résumé :
Frein majeur à l'inclusion dans les essais cliniques, la pré-sélection des sujets éligibles par la fouille manuelle des dossiers des patients est chronophage, mais peut être accélérée grâce à l'exploitation des dossiers informatisés. Augmenter l'efficacité et l'exhaustivité de l'identification des patients potentiellement éligibles à l'entrée dans les essais en pneumo-oncologie est possible en développant un outil de traitement automatique du langage pour vérifier les critères d'inclusion.

Pour chacun des huit critères d'inclusion aux essais cliniques étudiés, des algorithmes basés sur des expressions régulières ont été implémentés et évaluées. L'appariement sujet-protocole a été estimé entre les comptes rendus de Réunion de Concertation Pluri-professionnelles (RCP) stockés dans l'entrepôt de données hospitalières du CHRU de Tours et les protocoles publiés dans le registre français de l'institut contre le cancer (RECF) concernant les essais thérapeutiques ouverts à l'inclusion dans le Grand Ouest.

Ont été extraits 368 comptes rendus de RCP entre janvier et juillet 2021 ainsi que 16 protocoles d'essais ouverts en juillet 2021. Les performances pour détecter les critères d'inclusion dans les essais cliniques s'élevaient à 86 % de précision et 89 % de rappel en moyenne par protocole. L'outil appariait au moins un protocole à près d'un tiers des sujets (29,6 %) et identifiait au moins un sujet éligible pour la majorité des protocoles (81,3 %).

La pré-sélection automatisée des patients de pneumo-oncologie diminuerait la charge de travail avec une performance au moins aussi élevée que celle retrouvée dans la littérature. Les comptes rendus de RCP étant plus structurés que les fiches des protocoles, l'extraction des caractéristiques du patient montrait des performances plus élevées que celle des critères d'éligibilité. La précision de l'appariement sujet-protocole pourrait être augmentée en utilisant des variables de biologie. Tester l'outil sur les comptes rendus de RCP d'autres centres serait nécessaire pour estimer sa validité externe et permettre une interopérabilité des entrepôts de données pour des recrutements multicentriques, notamment avec la disponibilité du Ouest Data Hub.

Mots clés : "Tumeurs pulmonaires/statistiques et données numériques", "Stades des tumeurs/utilisation thérapeutique", "Réunion de concertation multidisciplinaire", "Traitement automatique du langage naturel", "Sélection du traitement du patient".

Jury :

Président du Jury :	Professeur Emmanuel RUSCH
Directeur de thèse :	Docteur Leslie GUILLON-GRAMMATICO
Membres du Jury :	Professeur Claude LINASSIER, Docteur Chloé PLICHON
Date de soutenance :	20 Octobre 2021