Problématique des entrepôts de données textuelles : dr Warehouse et la recherche translationnelle sur les maladies rares

Nicolas Garcelon

UNIVERSITE PARIS DESCARTES
ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS :
EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE
INSERM 1138, équipe 22

Problématique des entrepôts de données textuelles : Dr Warehouse et la recherche translationnelle sur les maladies rares

Par Nicolas Garcelon

Thèse de doctorat
Spécialité : Informatique biomédicale

Co Dirigée par Pr Anita Burgun et Pr Arnold Munnich

Présentée et soutenue publiquement le 29 novembre 2017

Devant le jury composé de

M Pascal STACCINI	Professeur	Rapporteur
M Patrick RUCH	Professeur	Rapporteur
Mme Anita BURGUN	Professeur	Directrice
M Arnold MUNNICH	Professeur	Directeur
Mme Natalia GRABAR	Professeur	Examinatrice
M Stanislas LYONNET	Professeur	Examinateur
Mme Brigitte SEROUSSI	Maitre de conférences	Examinatrice

« Si tu sais que c'est là une main, alors nous t'accordons tout le reste »

Ludwig Wittgenstein, *De la certitude*, 1.

C'est dans les mots que nous pensons. Nous n'avons conscience de nos pensées déterminées et réelles que lorsque nous leur donnons la forme objective, que nous les différencions de notre intériorité et par suite nous les marquons d'une forme externe, mais d'une forme qui contient aussi le caractère de l'activité interne la plus haute. C'est le son articulé, le mot, qui seul nous offre une existence où l'externe et l'interne sont si intimement unis. Par conséquent, vouloir penser sans les mots, c'est une tentative insensée. Et il est également absurde de considérer comme un désavantage et comme un défaut de la pensée cette nécessité qui lie celle-ci au mot. On croit ordinairement, il est vrai, que ce qu'il y a de plus haut, c'est l'ineffable. Mais c'est là une opinion superficielle et sans fondement ; car, en réalité, l'ineffable, c'est la pensée obscure, la pensée à l'état de fermentation, et qui ne devient claire que lorsqu'elle trouve le mot. Ainsi le mot donne à la pensée son existence la plus haute et la plus vraie.

Hegel, *Philosophie de l'esprit* (1817).

Remerciements

Je tiens à remercier tout d'abord, ma directrice, le Pr Anita Burgun, pour sa confiance et ses encouragements permanents après 16 ans de collaboration, ainsi que mon co-directeur le Pr Arnold Munnich qui m'a permis d'intégrer ce formidable projet qu'est l'Institut *Imagine*. Je remercie très vivement les professeurs Pascal Staccini et Patrick Ruch d'avoir accepté d'être rapporteurs de ce document. Merci de l'intérêt que vous avez porté à mon travail.
Je tiens à remercier le Pr Stanislas Lyonnet, le Pr Natalia Grabar et le Dr Brigitte Seroussi d'avoir accepté de participer au Jury de soutenance.

Je remercie particulièrement le Pr Rémi Salomon qui a porté avec moi ce projet d'entrepôt de données en texte libre à l'APHP depuis 6 ans et qui a été un soutien indéfectible pour son développement sur l'hôpital Necker-Enfants malades.

Je remercie chaleureusement M Guillaume Huart qui m'a permis de préparer cette thèse tout en continuant ma mission de responsable de la plateforme data science au sein de l'institut *Imagine*. Et je remercie l'actuelle direction d'*Imagine*, Mme Laure Boquet et Pr Stanislas Lyonnet, qui ont continué à me faire confiance pour mener de front ces deux activités.

J'exprime ma gratitude aux chercheurs et médecins de l'hôpital Necker et de l'Institut *Imagine* qui m'ont fait confiance, en particulier, le Pr Alain Fischer, le Pr Jeanne Amiel, le Pr Nadia Bahi-Buisson, le Dr Olivia Boyer, le Pr Smail Hadj-Rabia, le Pr Olivier Hermine, M Sven Kracker, le Pr Capucine Picard, le Pr Felipe Suarez, le Pr Rima Nabbout et beaucoup d'autres.

Je remercie le conseil scientifique « entrepôt de données » de Necker pour avoir suivi avec intérêt l'évolution du projet : Dr Célia Crétolle, Dr Marie-Alexandra Alyanakian, Pr Marie-France Mamzer, Mme Elisabeth Hulier-Ammar, Dr Nizar Mahlaoui.

Je remercie les membres du service informatique de Necker qui m'ont accueilli pendant 2 ans, et m'ont permis de rapidement intégrer les problématiques du système d'information de Necker.

L'ensemble de l'équipe qui m'a aidé dans ce travail. Vincent Benoit pour son soutien quotidien, Antoine Neuraz pour ses conseils judicieux, Bastien Rance pour ses conseils et sa capacité à relativiser, Hassan Faour pour son aide, Arthur Delapalme pour son énergie à revendre.

Je remercie le 4ieme étage de l'institut *Imagine* pour les déjeuners récréatifs, en particulier, Jeanne T, Isabelle B, Nicholas R, Nicolas D, Jérôme J, Pierre-Alexis M, Amélie C, Wilfried L, Romain M, Sabrina M.
Je remercie tous ceux avec qui j'ai échangé pendant ces 4 années de thèse et qui m'auront chacun permis à leur manière d'avancer.

Je remercie mes proches pour leur soutien et encouragement.

Publications

B. Campillo-Gimenez, N. Garcelon, P. Jarno, J.M. Chapplain, M. Cuggia, Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France, Stud Health Technol Inform. 192 (2013) 572–575.

K. Bouchireb, O. Boyer, O. Gribouval, F. Nevo, E. Huynh-Cong, V. Morinière, R. Campait, E. Ars, D. Brackman, J. Dantal, P. Eckart, M. Gigante, B.S. Lipska, A. Liutkus, A. Megarbane, N. Mohsin, F. Ozaltin, M.A. Saleem, F. Schaefer, K. Soulami, R. Torra, N. Garcelon, G. Mollet, K. Dahan, C. Antignac, NPHS2 mutations in steroid-resistant nephrotic syndrome: a mutation update and the associated phenotypic spectrum, Hum. Mutat. 35 (2014) 178–186. doi:10.1002/humu.22485.

N. Garcelon, V. Courteille, A. Fischer, N. Mahlaoui, Epidemiology of PID: Innovative New Way to Identify Patients in the CEREDIH Registry Through a Medical Data Warehouse, J. Clin. Immunol. 34 (2014) S361–S362.

N. Garcelon, A. Neuraz, V. Benoit, R. Salomon, A. Burgun, Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse, J Am Med Inform Assoc. (2016). doi:10.1093/jamia/ocw144.

N. Garcelon, A. Neuraz, V. Benoit, R. Salomon, S. Kracker, F. Suarez, N. Bahi-Buisson, S. Hadj-Rabia, A. Fischer, A. Munnich, A. Burgun, Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack, Journal of Biomedical Informatics. 73 (2017) 51–61. doi:10.1016/j.jbi.2017.07.016.

N. Garcelon, A. Neuraz, R. Salomon, H. Faour, V. Benoit, A. Delapalme, A. Munnich, A. Burgun, B. Rance, A clinician friendly data warehouse oriented toward narrative reports: Dr Warehouse, Journal of Biomedical Informatics. Submitted in September 2017.

N. Garcelon, A. Neuraz, R. Salomon, V. Benoit, N. Bahi-Buisson, A. Munnich, A. Burgun, B. Rance, Next Generation Phenotyping using narrative reports in a rare disease Clinical Data Warehouse, Orphanet Journal of Rare Diseases. Submitted in September 2017.

Communications

N. Garcelon, Des données médicales à la connaissance : entrepôts et fouilles de données, Ann Dermatol Venereol. 142 (2015) S389-390. doi:10.1016/j.annder.2015.10.171.

Garcelon N. Plenary Lecture: From data to knowledge: biomedical data warehouse and data mining. 13th Congress of the European Society for Pediatric Dermatology (ESPD). Paris, 28th May 2016

Garcelon N. Les entrepôts de données et le big data au service de la médecine. SUN 2017. Palais des congrès. January, 26 2017.

Garcelon N. Dr Warehouse, un entrepôt de données orienté maladie rare. Journée Maladie Rare. Institut Imagine. October 11 2016.

Garcelon N. The production of healthcare data: ensuring that the system starts and ends with the patient. Healthcare Data Institute. April 19, 2017.

Garcelon N. Journée sur la recherche en santé dans ses aspects éthiques et réglementaires (données, algorithmes). Débat public initié par la CNIL sur les enjeux éthiques et de société soulevés par les algorithmes et l'Intelligence Artificielle (IA). Paris, 15 Septembre 2017

Publications grand public

Garcelon N. Interview for a publication in l'Express: Les États-Unis font main basse sur la médecine de demain. Stéphanie Benz, published October 14, 2016 - L'express

Garcelon N. Editorial for the Healthcare Data Institute : The production of healthcare data: ensuring that the system starts and ends with the patient. April, 19 2017. [http://healthcaredatainstitute.com/2017/04/19/the-production-of-healthcare-data-ensuring-that-the-system-starts-and-ends-with-the-patient/]

Dépôt de logiciels à l'Agence de Protection des Programmes

Dr Warehouse – Un entrepôt de données orienté document – licence open source GNU GPL v3

BioPancarte – Un logiciel web hospitalier pour personnaliser la visualisation des résultats biologiques

Gecko – Un logiciel web hospitalier pour la gestion des études cliniques : pré inclusion et inclusion des patients, consentement, rendez vous, suivi des patients, tableau de bord.

Auxo – Un logiciel web hospitalier pour créer, visualiser et partager des courbes de croissances (poids, taille, périmètre crânien).

Agrément éthique

Nous avons obtenu un avis favorable du Comité de Protection des Personnes Ile de France II (IRB registration number 00001072) enregistré sous la référence 2016-06-01.

Résumé

La réutilisation des données de soins pour la recherche s'est largement répandue avec le développement d'entrepôts de données cliniques. Ces entrepôts de données sont modélisés pour intégrer et explorer des données structurées liées à des thesaurus. Ces données proviennent principalement d'automates (biologie, génétique, cardiologie, etc) mais aussi de formulaires de données structurées saisies manuellement.

La production de soins est aussi largement pourvoyeuse de données textuelles provenant des comptes rendus hospitaliers (hospitalisation, opératoire, imagerie, anatomopathologie etc.), des zones de texte libre dans les formulaires électroniques. Cette masse de données, peu ou pas utilisée par les entrepôts classiques, est une source d'information indispensable dans le contexte des maladies rares. En effet, le texte libre permet de décrire le tableau clinique d'un patient avec davantage de précisions et en exprimant l'absence de signes et l'incertitude. Particulièrement pour les patients encore non diagnostiqués, le médecin décrit l'histoire médicale du patient en dehors de tout cadre nosologique. Cette richesse d'information fait du texte clinique une source précieuse pour la recherche translationnelle. Cela nécessite toutefois des algorithmes et des outils adaptés pour en permettre une réutilisation optimisée par les médecins et les chercheurs.

Nous présentons dans cette thèse l'entrepôt de données centré sur le document clinique, que nous avons modélisé, implémenté et évalué. A travers trois cas d'usage pour la recherche translationnelle dans le contexte des maladies rares, nous avons tenté d'adresser les problématiques inhérentes aux données textuelles: (i) le recrutement de patients à travers un moteur de recherche adapté aux données textuelles (traitement de la négation et des antécédents familiaux), (ii) le phénotypage automatisé à partir des données textuelles et (iii) l'aide au diagnostic par similarité entre patients basés sur le phénotypage.

Nous avons pu évaluer ces méthodes sur l'entrepôt de données de Necker-Enfants Malades créé et alimenté pendant cette thèse, intégrant environ 490 000 patients et 4 millions de comptes rendus. Ces méthodes et algorithmes ont été intégrés dans le logiciel Dr Warehouse développé pendant la thèse et diffusé en Open source depuis septembre 2017.

Mots clés : Entrepôt de données, Fouille de données, Maladies rares, Phénotypage, Recherche d'information

Abstract

The repurposing of clinical data for research has become widespread with the development of clinical data warehouses. These data warehouses are modeled to integrate and explore structured data related to thesauri. These data come mainly from machine (biology, genetics, cardiology, etc.) but also from manual data input forms.

The production of care is also largely providing textual data from hospital reports (hospitalization, surgery, imaging, anatomopathologic etc.), free text areas in electronic forms. This mass of data, little used by conventional warehouses, is an indispensable source of information in the context of rare diseases. Indeed, the free text makes it possible to describe the clinical picture of a patient with more precision and expressing the absence of signs and uncertainty. Particularly for patients still undiagnosed, the doctor describes the patient's medical history outside any nosological framework. This wealth of information makes clinical text a valuable source for translational research. However, this requires appropriate algorithms and tools to enable optimized re-use by doctors and researchers.

We present in this thesis the data warehouse centered on the clinical document, which we have modeled, implemented and evaluated. In three cases of use for translational research in the context of rare diseases, we attempted to address the problems inherent in textual data: (i) recruitment of patients through a search engine adapted to textual (data negation and family history detection), (ii) automated phenotyping from textual data, and (iii) diagnosis by similarity between patients based on phenotyping.

We were able to evaluate these methods on the data warehouse of Necker-*Enfants Malades* created and fed during this thesis, integrating about 490,000 patients and 4 million reports. These methods and algorithms were integrated into the software Dr Warehouse developed during the thesis and distributed in Open source since September 2017.

Keywords: Data warehouse, Data mining, Rare diseases, Phenotyping, Information retrieval

Table des matières

Abréviations

APDS	Syndrome PI3K Delta Activé
AQL	Archetype Querying Language
ARN	Acide ribonucléique
ASCII	American Standard Code for Information Interchange
CHU	Centre Hospitalier Universitaire
CIM10	Classification Internationale des Maladies version 10
CNIL	Commission nationale de l'informatique et des libertés
CUI	Concept Unique Identifier
DGOS	Direction Générale de l'Organisation des Soins
DPI	Dossier Patient Informatisé
DRG	Diagnostic Related Group
EAV	Entity-Attribute-value
EMERSE	Electronic Medical Record Search Engine
ETL	Extract, Tranform and Load
FHIR	Fast Healthcare Interoperability Resources
FTP	File Transfer Protocol
GUAM	Generalized Update Access Method
HL7	Health Level 7
HPO	Human Phenotype Ontology
HTML	HyperText Markup Language
i2b2	Informatics for Integrating Biology & the Bedside
IMS	Information Management System
LOINC	Logical Observation Identifiers Names and Codes
MAP	Mean Average Precision
MIT	Master of Information Technology
NER	Named Entity Recognizer
OLAP	OnLine Analytical Processing
OMIM	Online Mendelian Inheritance in Man
PDF	Portable Document Format
PMSI	Plan de Médicalisation du Système d'Information
RIM	Reference Information Model
RUM	Résumé d'Unité Médicale
SGBD	Système de Gestion des Bases de Données
SHRINE	The Shared Health Research Information Network
SQL	Structured Query Language
TF-IDF	Term Frequency-Inverse Document Frequency
UMLS	Unified Medical Language System
XML	Extensible Markup Language

1. Introduction

Il faut remonter jusqu'en 1600 avant J.-C. pour trouver les premiers rapports médicaux rédigés sur un papyrus égyptien (Al-Awqati, 2006). Il s'agit des premières études de cas à visée pédagogique. A partir du 17ième siècle, le processus s'amplifie toujours dans un objectif d'enseignement mais aussi afin de réaliser des corrélations anatomiques et diagnostiques (Gillum, 2013). Il faut attendre le début du 19ième siècle à Paris et Berlin pour voir l'apparition du premier dossier patient sous forme de feuilles volantes (Hess, 2010). En France, l'émergence dans les années 1920 d'une classe moyenne qui refuse de se faire soigner dans des hôpitaux réservés aux pauvres et indigents, mais qui n'a pas suffisamment de revenus pour aller dans des cliniques privées, va initier dans les années 50 une « humanisation » de l'hôpital et remettre le patient au centre du soin (Nardin, 2010) et donc profondément modifier la manière de recueillir les données. A partir des années 70, le dossier patient se structure et s'archive sous forme de dossier infirmier, de comptes rendus médicaux, de lettres de sortie. A la fin des années 90, les comptes rendus s'informatisent et se structurent davantage. L'informatisation s'accélère notamment pour évaluer les coûts et les remboursements des soins (PMSI, DRG) (*Ordonnance no 96-346 du 24 avril 1996 portant réforme de l'hospitalisation publique et privée*, n.d.). A partir des années 2000, le dossier patient informatisé (DPI) commence à faire son apparition dans les hôpitaux pour se généraliser en 2010.

Avec l'informatisation du dossier patient, la réutilisation de ces données pour la recherche, pour le management hospitalier et pour l'enseignement devient de plus en plus immédiate, au point que nous structurons les DPIs non plus pour améliorer la prise en charge du patient, mais pour faciliter cette réutilisation. Cette dérive du DPI rend son remplissage laborieux et finalement peu efficace. En effet, les médecins préfèrent utiliser les champs en texte libre plutôt que de cocher des cases. Le texte libre leur permet de mieux préciser leur pensée, d'y inscrire des doutes et des absences de signe, des hypothèses diagnostics (Hanauer et al., 2015; Raghavan et al., 2014; Shivade et al., 2014). Dans le contexte des maladies rares, le texte libre reste le moyen idéal de conserver la richesse phénotypique du patient tout en conservant la relation malade / médecin autour d'une médecine narrative (Charon, 2012). Par ailleurs, certains patients restent sans diagnostic au début de leur prise en charge, le médecin doit donc pouvoir décrire les signes cliniques du patient sans a priori sur son approche sémiologique afin de documenter l'état clinique du patient le plus précisément possible.

Les équipes d'informatique médicale ou « data scientists » doivent développer des méthodes permettant de réutiliser les données produites dans le cadre du soin pour la recherche et l'enseignement sans dénaturer leur premier objectif: soigner les malades (Rosenbloom et al., 2011). Toll illustre dans un article ce paradoxe par un dessin fait par une patiente de 7 ans, Figure 1 (Toll, 2012).

Figure 1: Dessin d'une patiente de 7 ans illustrant la consultation avec le médecin. Le dessin représente l'enfant sur la table de consultation, sa mère, sa sœur ainsi que le petit frère. Le médecin tourne le dos à la patiente pour saisir les données sur l'ordinateur. Tout le monde est souriant.

Comment réutiliser le dossier de soin ?

Le système d'information hospitalier reste encore très hétérogène et cloisonné d'un logiciel de soin à l'autre. Les hôpitaux mettent en place depuis une dizaine d'années le DPI afin d'uniformiser leur parc logiciel et ainsi réduire le nombre de sources de données de production. Malgré cela, certaines spécialités médicales ont leur propre logiciel, la réanimation, l'anesthésie, l'imagerie, la génétique, l'imagerie anténatale (qui nécessite un identifiant particulier pour les fœtus), le PMSI, les urgences etc. Ces spécialités sont de plus en plus intégrées dans le DPI pour rationnaliser les coûts de maintenance, souvent au détriment des fonctionnalités nécessaires à prendre en compte leurs spécificités. Certaines spécialités résistent à cette intégration afin de garder une souplesse dans les outils qu'ils utilisent. Afin de pallier cette dispersion des données qui peut rendre compliquée la prise en charge des patients, les DPIs peuvent fonctionner comme « data repository ». Cela permet de stocker dans le DPI les données produites dans d'autres logiciels. La DGOS publie annuellement un état des lieux de l'informatisation des établissements de soin (DGOS, 2017). Cet atlas ne prend en compte que

les logiciels dits institutionnels et témoigne déjà de la grande hétérogénéité du parc logiciel hospitalier.

Au delà de la diversité des logiciels qui coexistent dans un hôpital, les données médicales sont aussi contenues dans des logiciels obsolètes inutilisés dont les données n'ont pas été reprises dans un logiciel de remplacement. Suivant l'année de mise en place du DPI, le nombre de données ainsi « perdues » est difficilement évaluable sans faire le travail de les récupérer, mais on peut le supposer élevé.

Dans le contexte des maladies rares, toutes les données sont importantes en raison, par définition, de la rareté des patients atteints. Il s'agit de pouvoir réutiliser aussi bien les données des logiciels institutionnels que des archives, des logiciels utilisés par une seule équipe, des bases de données réalisées dans le cadre d'un suivi de cohorte, des bases de données « maison » non maintenues etc.

Les données dans ces logiciels sont stockées dans des formats très hétérogènes :

- données codées par un thésaurus local ou national, (résultat biologique, PMSI)
- données structurées non codées (par exemple « diabete : oui », « taille : 180cm »)
- données textuelles semi structurées (texte divisé en section « motif », « conclusion », « résultat » etc.)
- données textuelles non structurées (comptes rendus au format Word, PDF, ASCII, HTML)

Au sein d'un même logiciel, tous ces formats peuvent coexister.

Réutiliser ces données nécessite de pouvoir y accéder rapidement sans compromettre le logiciel de production. L'entrepôt de données biomédical permet de répondre à cette problématique en colligeant toutes les données en une seule base de données par la copie régulière du contenu des bases de production de soins. Il existe autant d'entrepôts de données qu'il y a de cas d'usage de celui-ci. Il s'agit donc de proposer un entrepôt de données qui réponde à notre vision de la réutilisation des données de soins. Nous avons déjà évoqué précédemment la nécessité de privilégier la relation médecin patient au cœur du système d'information hospitalier. Pour cela, l'entrepôt de données doit intégrer et gérer les données textuelles. Cela implique des outils adaptés et cela va impacter la structure et les fonctionnalités de l'entrepôt de données.

Nous avons considéré trois cas d'usage de cet entrepôt de données :

- La recherche de patients
- La fouille de données
- L'aide au diagnostic

La recherche de patients

C'est la fonctionnalité la plus fréquente dans les entrepôts de données biomédicaux (Shin et al., 2014). Il s'agit ici de donner aux cliniciens la possibilité de rechercher des patients à partir de

leurs données cliniques. Il nous faut définir sur quels critères le clinicien peut effectuer la recherche (données codées, données textuelles, données démographiques). Il nous faut aussi définir quelle est la finalité de ce moteur de recherche. Suivant les résultats attendus, le modèle de données, le moteur, l'interface d'interrogation et de visualisation ainsi que les fonctionnalités seront différents. Trois cas d'usage peuvent être définis autour de la recherche d'information dans les entrepôts de données:

- réaliser des études de faisabilité (Deshmukh et al., 2009; McCowan et al., 2015). Le moteur renvoie un nombre de patients répondant aux critères de recherche.
- recruter des patients dans des études (Brooks et al., 2009; Murphy et al., 2000, 1999; Weng et al., 2010). Cela nécessite des données détaillées et nominatives ou ré-identifiables pour pré-screener les patients et éventuellement les contacter.
- retrouver des patients pour aider à prendre une décision sur un nouveau patient. Ce type de finalité peut être en complément des réunions de concertation pluridisciplinaire. Frankovitch a montré l'intérêt d'un entrepôt de données pour choisir le meilleur traitement pour un patient atypique à partir de patients présentant les mêmes caractéristiques dans l'entrepôt de données (Frankovich et al., 2011).

Il s'agit aussi de définir les stratégies à mettre en œuvre afin d'améliorer les performances du moteur de recherche autour de la recherche textuelle. Les performances sont améliorées en réduisant le bruit et le silence. Le bruit se définit par les patients faux positifs, c'est à dire les patients retrouvés par le moteur mais ne correspondant pas aux critères de recherche. Le silence se définit par les patients faux négatifs, c'est à dire les patients correspondants aux critères de recherche mais non retrouvés par le moteur.

Si les comptes rendus médicaux décrivent les signes cliniques des patients, ils font aussi état de l'absence de certains signes permettant d'exclure un diagnostic. Ils décrivent aussi les antécédents familiaux du patient, particulièrement lorsqu'il s'agit d'un enfant atteint d'une maladie génétique. Il est donc indispensable de développer des stratégies de détection de la négation et des antécédents familiaux afin de réduire le nombre de patients faux positifs dans les résultats du moteur de recherche.

Pour réduire le silence (ou le nombre de patients faux négatifs), une stratégie d'enrichissement terminologique étend la requête aux synonymes et hyponymes des termes recherchés.

La fouille de données

Le deuxième cas d'usage attendu est la possibilité d'établir une description phénotypique d'une population. Au delà des caractéristiques démographiques telles que l'âge, le sexe, la géolocalisation des patients, il s'agit de proposer des méthodes permettant de détecter des nouvelles associations phénotypiques à partir des données hospitalières (Hripcsak and Albers, 2013).

Dans un premier temps, Il s'agira d'extraire depuis les comptes rendus cliniques les concepts médicaux puis de les trier par pertinence pour une sous population définie par des critères diagnostiques ou génétiques.

La recherche translationnelle : retrouver des patients similaires

Dans l'écosystème de l'institut Imagine et de l'hôpital Necker Enfants Malades, de nouvelles mutations génétiques délétères sont identifiés chaque année par les équipes de recherche. La requête PubMed « gene mutation and (*imagine* institute or institut imagine or Necker hospital) » renvoie 114 publications en 2015, 101 en 2016 et 61 pour le 1er semestre 2017. L'objectif est de proposer une méthode simple permettant de retrouver des patients non diagnostiqués déjà venus à l'hôpital Necker à partir des données du patient index diagnostiqué. Ces patients seraient alors éligibles pour être screenés sur le gène causal nouvellement identifié.

Le moteur de recherche permet de retrouver des patients répondant à des critères de recherche booléens. Il est toutefois difficile de définir ces critères à partir d'un seul patient diagnostiqué. Retrouver des patients par similarité permet de prendre en compte toute la complexité phénotypique décrite dans les comptes rendus, sans a priori sur la pertinence de garder tel ou tel signe clinique. Dés 2009 Ruch souligne l'importance d'adresser les défis liés à l'exploitation des données structurées et non structurées pour les outils d'aide à la décision basés sur les données cliniques (Ruch, 2009).

Les maladies rares

L'hôpital Necker-Enfants Malades est un hôpital de 600 lits. Il regroupe un tiers de l'activité de pédiatrie des Hôpitaux de Paris-Assistance Publique et presque la moitié de la chirurgie pédiatrique. A partir des années 1950, l'hôpital se spécialise sur les maladies génétiques. Il héberge 32 centres maladies rares, et coordonne 15 centres de référence maladie rare ("Introducing the Necker-Enfants Malades hospital," 2015).

L'institut *Imagine* est créé en 2007, le bâtiment est inauguré en 2014 sur le campus de l'hôpital Necker. Il rassemble plus de 900 chercheurs, médecins et personnels de santé avec pour objectifs d'accélérer les synergies, favoriser le transfert des connaissances et ainsi, trouver plus vite de nouveaux traitements et diagnostics attendus par les patients et leurs familles. L'institut *Imagine* est un lieu de recherche translationnelle à l'interface entre l'hôpital et la recherche, dans lequel les patients sont suivis par les médecins de Necker et diagnostiqués par les équipes de recherche de l'*institut* ("L'Institut Imagine," n.d.).

Les données cliniques produites au cours du soin ainsi que les bases de données produites par les équipes de recherche sont une richesse unique en France de par la prévalence élevée de maladies rares ou non encore diagnostiquées.

2. Etat de l'art

La notion de business intelligence (BI) est décrite pour la première fois en 1959 par Hans Peter Luhn, ingénieur à IBM (Heinze, 2014). Cette théorie est ensuite recontextualisée dans le domaine informatique par Dresner en 1989. Il parle déjà de modèle à base de faits. A partir du début des années 1990, le secteur économique et industriel réclame des analyses de plus en plus poussée de leur activité au travers de leurs outils de production. Il devient alors nécessaire de proposer de nouveaux modèles dédiés au stockage et à l'analyse de gros volumes de données.
Bill Inmon, considéré comme le père des entrepôts de données, définit l'entrepôt de données en 1992 comme « une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision » (Inmon, 1992). Il ajoute 17 ans plus tard : « you do not buy a technology and have a data warehouse. Instead, you design and build the proper structure, and then you seek out the best technology to help you access and analyze the data. Most vendors that offer to sell you a data warehouse are pulling your leg. » (Inmon, 2009)
Depuis les années 1960, les données informatiques n'ont cessé d'augmenter en volume et de se diversifier dans leur format (texte, image, vidéo, son, génome, etc.). Les avancées technologiques ont permis d'augmenter les capacités de stockage et la puissance de calcul des systèmes informatiques. Les entrepôts de données doivent faire face à de nouveaux besoins, de nouveaux défis notamment dans le contexte médical actuel de médecine personnalisée. Les entrepôts de données se définissent par leur modèle, les données chargées et les fonctionnalités qui leurs sont associées.

2.1. Sélection d'entrepôts de données

Nous avons réalisé un état de l'art par l'analyse de 16 entrepôts de données publiés. Nous les avons sélectionnés par une recherche sur PubMed. Les critères de sélection sont:

- il doit contenir des données cliniques de patients. Nous avons exclu les entrepôts de données documentaires
- il doit proposer une solution originale. Nous avons exclu les entrepôts de données qui réutilisaient un modèle existant (par exemple i2b2).
- la publication de référence doit donner suffisamment d'information sur les aspects techniques de l'entrepôt de données.

Les entrepôts de données retenus sont :

- **CATCH/IT** (Berndt et al., 1998) développé par le College of Public Health University of South Florida
- **Data-Warehouse-Concept for Clinical and Biological Information** (Brammen et al., 2005) développé par le CHU de Giessen en Allemagne

- **I2B2** (Murphy et al., 2006) développé par Harvard medical school / MIT
- **EMERSE** (Hanauer, 2006) développé par le Michigan medical school
- **Radbank** (Rubin and Desser, 2008) développé par le Stanford Medical Informatics
- **STRIDE** (Lowe et al., 2009) développé par l'université de Stanford
- **Enterprise Data Trust at Mayo Clinic** (Chute et al., 2010) développé par la Mayo Clinic
- **NC CATCH** (Studnicki et al., 2010) développé par l'université de Caroline du Nord
- **DW4TR** (Hu et al., 2011) développé par Windber Research Institute (USA)
- **ERS Kyoto** (Yamamoto et al., 2012) développé par Department of Clinical Trial Design and Management, Translational Research Centre, Kyoto University Hospital, Kyoto, Japan
- **VANDERBILT DWH** (Danciu et al., 2014) développé par Vanderbilt Institute for Clinical and Translational Research
- **Archetype-based DWH** (Marco-Ruiz et al., 2015) développé par Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Norway
- **METEOR** (Puppala et al., 2015) par le Houston Methodist Hospital Research Institute
- **Starmaker** (Krasowski et al., 2015) développé par University of Iowa Hospitals and Clinics
- **Consore** (Heudel et al., 2016) par Sword (France)
- **SMEYEDAT** (Kortüm et al., 2017) par University Eye Hospital, Ludwig-Maximilians-University Munich, Munich, Germany.

Le Tableau 1 synthétise cette analyse.

Tableau 1 : Entrepôts de données biomédicaux

Nom	PMID	Modèle base de données	Source de données	Données structurées	Texte libres	De-identification	Moteur de recherche en texte libre	Technologie texte libre	Interface adaptée au texte libre	Autres Fonctionnalités	SGBD	Partagé
CATCH/IT:	9929220	Schéma en étoile	226 indicateurs de santé depuis 12 comtés de Floride	Oui	Non	Anonymisation	Non	-	-	Renvoie données agrégées	Oracle	Non
Data-Warehouse-Concept for Clinical and Biological Information	16160228	modèle EAV + relationnel + sémantique	DPI + micro array	Oui	Non ?	de-identification	Non	-	-	Création de datamart	Oracle	Non
I2B2	17238659	Schéma en étoile	Potentiellement tout	Oui	Possibilité d'enregistrer le texte libre, mais extraction des concepts dans la plupart des cas	de-identification	Non	SQL	-	Comparaison de cohortes, développement de plugins par la communauté	Oracle / postgres	Open source
EMERSE	17238560	document repository	Compte rendu du DPI (format HL7)	Non	Oui	de-identification	Oui	Apache Lucene + solr	Oui	Partage de requêtes. Exploration du dossier d'un patient. Possibilité d'exporter des cohortes vers i2b2	Oracle	Free for academic
Radbank	18312970	XML	Démographique, CR radiologie, CR anapath	Oui	Oui	nominative	Oui, par des requêtes SQL manuelles. Possibilité de cibler des sections. Pas de prise en compte de la négation	Module Oracle text	Non		Oracle 9	Non
STRIDE	20351886 2009	Entité attribut	DPI HL7, Base de données recherche, biobank, démographique	Oui	Oui	Nominative + de-identification	Oui, par des requêtes SQL manuelles. Pas de prise en compte de la négation ou des antécédents familiaux	Module Oracle text	Non	interface pour inclure / exclure les patients d'une cohorte. Suite logiciel pour la recherche translationnelle	Oracle 11g	Non

Nom	PMID	Modèle base de données	Source de données	Données structurées	Texte libres	De-identification	Moteur de recherche en texte libre	Technologie texte libre	Interface adaptée au texte libre	Autres Fonctionnalités	SGBD	Partagé
Enterprise Data Trust at Mayo Clinic (EDT-MC)	20190054 2010	Orienté Objet	Démographique, Séjour, matériel biologique, diagnostics, maladies, localisation, prescription, HL7	Oui	Texte libre transformé en données codées	de-identification	Non	-	-		IBM	Non
NC CATCH	23569592 2010	Multidimensionnel	données démographiques / démographiques au niveau des régions de recensement; mortalité, grossesses, naissances, courriers de sortie, visites de salles d'urgence, données sur les enquêtes sur les facteurs de risque comportementaux, incidence du cancer et données sur le traitement, données sociales, économiques et sanitaires.	Oui	Non	de-identification	Non	-	-	Accès anonyme aux données agrégées via interface multidimensionnelle de type hyper cube	?	Non
DW4TR	21872681 2011	Entité attribut	Questionnaires de recherche, données génétiques, données chirurgicales, psychologie, antécédents familiaux, pathologie, habitudes alimentaires	Oui	Non	de-identification	Non	-	-	ABB (navigateur multidimensionnel) et ISIV (Explorateur du dossier patient).	Oracle 10g	Non
ERS	23117567 2012	Relationnelle (en étoile détourné)	DPI (démographique, diagnostics, examens laboratoires, prescriptions, chirurgie)	Oui	Non	de-identification	Non	-	-	OLAP pour générer des requêtes SQL à travers une interface web	?	Non
VANDERBILT	24534443 2014	Non précisé - relationnel ?	DPI, PMSI, comptes rendus cliniques, toutes les données structurées (laboratoire etc.)	Oui	Oui	Nominative + de-identification	Mots clés et expressions régulières	?	?	Deux ED : RD (nominatif) et SD (déidentifié). Accès Web. Export possible vers RedCap.	IBM Netezza	Non

Nom	PMID	Modèle base de données	Source de données	Données structurées	Texte libres	De-identification	Moteur de recherche en texte libre	Technologie texte libre	Interface adaptée au texte libre	Autres Fonctionnalités	SGBD	Partagé
Archetype-based DWH	26094821 2014	Modèle Archetype	Résultats biologiques	Oui	Non	de-identification	Non	-	-	Utilisation de AQL pour définir des indicateurs de résultats biologiques	Think!EHR	Non
METEOR	26126271 2015	Schéma en étoile - basé sur i2b2	Cliniques, démographiques, génomiques, imagerie, administratives et économiques, et bases de données recherche, Anapath (biopsie des poumons)	Oui	Texte libre transformé en données codées	de-identification	Non	-	-	Readmit (score de risque de réadmission) + Motte (retrouver des indicateurs cliniques dans le texte libre)	Microsoft SQL	Non
Starmaker	26284156 2015	Relationnel	Parcours de soin, examens biologiques, prescription, administration des médicaments, allergie, immunisation, PMSI (diagnostics et actes)	Oui	Non	de-identification	Non	-	-	Recruter des patients et extraction de données cliniques. Export au format texte pour analyse par des logiciels statistiques	?	Commercial
Consore	27816168 2016	non précisé - multidimensionnel ?	Les données des centres anti cancéreux (DPI, biologie etc)	Oui	Texte libre transformé en données codées	de-identification	Non. Un champ en texte libre permet d'afficher des suggestions de termes en lien avec les données codées. Il ne s'agit pas à proprement parler de recherche en texte libre	Non précisé	?	Création de panier de patients, requêtage multi centrique, partage de requête	Park Street Solutions	Commercial
SMEYEDAT	28365240 2017	Schéma relationnel, centré sur le patient.	PMSI, DPI, Démographique, mesures de l'oeil	Oui	Oui	Nominative	Oui	Qlikview	-	sélection de patients, statistiques, export des données : spécifique à l'ophtalmologie	Microsoft SQL + Qlikview	Non

2.2. Les modèles de données

2.2.1. Les types de modèle

Jusqu'aux années 1960, les données étaient stockées dans des fichiers dits « file-based system ». La mission Apollo de 1963 nécessita la création d'un nouveau système permettant de stocker un grand nombre de données et de représenter des relations plus complexes (Connolly and Begg, 2004; Sumathi and Esakkirajan, 2007). La North American Aviation (NAA, maintenant Rockwell International) développe alors le GUAM (Generalized Update Access Method) qui est la première base de données basée sur un modèle hiérarchique. IBM rejoint la NAA au milieu des années 1960, pour développer et intégrer le GUAM dans l'IMS (Information Management System) encore utilisé aujourd'hui. Parallèlement à cette initiative, Charles Bachmann de la General Electric développe une base de données basée sur un modèle en réseau (Network DBMS). Ce modèle fera école et entrainera les premiers travaux de normalisation des définitions de modèle de données. Ces modèles en hiérarchie et en réseau restent néanmoins compliqués à manipuler et à maintenir. En 1970, Edgar Franck Codd, employé d'IBM, publie les fondements du modèle relationnel (Codd, 1970) qui ouvre le marché des bases de données. En 1979, IBM D2 et Oracle Database sont les premiers Systèmes de Gestion de Bases de Données (SGBD) relationnelles mis sur le marché.

Dans les années 1990, les modèles objets et objets-relations apparaissent. Si ces nouveaux modèles permettent de manipuler des données complexes et composites, ils ne percent que faiblement face aux géants que sont devenus les modèles SGBD Relationnelles.

A partir des années 2000, l'augmentation exponentielle des données produites par les marchés financiers, puis par les géants du Web, ont amené à une réflexion sur un nouveau modèle de SGBD non relationnel. Il s'agit alors de simplifier les échanges, d'améliorer les performances de calcul, et de rendre le modèle scalable. En 2009, le modèle NoSQL (Not only SQL) nait par opposition au modèle relationnel. Il regroupe en réalité plusieurs types de modèle non relationnel (orienté-agrégats, orienté-graphes, orienté-sans schéma) (Sadalage and Fowler, 2012).

2.2.2. Le modèle des entrepôts de données biomédicaux

Les entrepôts de données biomédicaux décrits dans la littérature ne fournissent que rarement leur schéma de modèle de données relationnel, car considéré comme spécifique à leur problématique et finalement peu généralisable. Il s'agit en général davantage de publier la faisabilité d'un tel entrepôt et les résultats obtenus que l'entrepôt en tant que tel.

Schéma en étoile

Conçu en 1993, ce modèle dimensionnel est une alternative au modèle classique entité-relation des SGBD relationnelle (Dinu and Nadkarni, 2007; Stead et al., 1983). La table de faits est la clé de voûte de ce modèle. Elle stocke tous les indicateurs de performance et représente le

centre d'un schéma dit « en étoile ». Les tables environnantes correspondent aux référentiels nécessaires à la description des faits. A l'origine, ce modèle a été pensé pour répondre à des questions économiques et industrielles.

En 2006, i2b2 est le premier entrepôt de données distribué gratuitement (Murphy et al., 2006). Cette volonté impacte la manière dont il crée le modèle de données, car l'objectif n'est plus une problématique locale, mais doit être généralisable. Le modèle proposé est un modèle dimensionnel en étoile inspiré des modèles classiques en informatique décisionnelle. La table des faits contient tous les évènements atomiques d'un patient. Un fait atomique est une donnée numérique ou textuelle reliée à un patient, un séjour, un praticien et à un référentiel (un thesaurus local ou national). Ce modèle ne suffisant pas à représenter la complexité des données médicales, il a été ajouté en 2010 la notion de modificateur (*modifiers*) et d'instance (*instance_num*) (Murphy, 2011a). Le modificateur précise l'expression d'un concept. Il est associé à un numéro d'instance qui permet de regrouper des faits partageant un même périmètre : par exemple, le fait « aspirine » sera répété autant de fois que nécessaire avec le même numéro d'instance afin de préciser les modificateurs « dose », « fréquence » et « mode d'administration ». Si ce modèle rend généralisable l'intégration d'une grande partie des données médicales, son interrogation peut devenir complexe et nécessite de connaître le référentiel et les modificateurs associés.

Chose étonnante, en 2009, Inmon considère qu'un modèle dimensionnel en étoile n'est pas un entrepôt de données (Inmon, 2009). Il ajoute que le modèle en étoile est trop dépendant de l'usage que l'on en fait et des exigences des utilisateurs. D'après lui le schéma relationnel entité-relation est le plus évolutif. C'est sans doute pourquoi le schéma en étoile est rarement utilisé dans sa version la plus stricte. Il est généralement utilisé comme élément central permettant de justifier une interopérabilité, mais il est rapidement adapté en ajoutant des tables pour répondre aux besoins spécifiques des utilisateurs, pour s'adapter à des types de données. Sengupta et al. ont ainsi développé un entrepôt de données en étoile en ajoutant une table spécifique pour traiter les données d'imagerie (Sengupta et al., 2013). L'entrepôt de données Meteor est une adaptation du modèle i2b2 (Puppala et al., 2015).

TranSMART réutilise le modèle i2b2 en intégrant des tables supplémentaires notamment pour héberger les données omics à haute dimension (Scheufele et al., 2014; Wang et al., 2014).

Modèle Entité-Valeur-Attribut (EAV) et modèle orient objet

En 2005, Brammen et al. développent un entrepôt de données contenant des données cliniques et biologiques (notamment les résultats des microarray sur ARN) (Brammen et al., 2005). Ils construisent leur entrepôt de données sur un schéma Entité-Attribut-Valeur basé sur HL7 RIM. Il est combiné à un modèle relationnel pour les données administratives.

Stride, développé par le Stanford University Medical Center (Lowe et al., 2009), est conçu sur un schéma Entité-Attribut-Valeur développé sur le SGBD Relationnelle Oracle. Il est conçu sur un model orienté objet. A partir des travaux de Lyman et al., ils ont mappé leurs données sur le référentiel HL7 Reference Information Model. L'équipe de la Mayo Clinic a aussi aligné son modèle de données sur ce référentiel en collaboration avec IBM (Chute et al., 2010).

Autres types de modèle

Radbank (Rubin and Desser, 2008) est développé sous Oracle, il s'agit d'un entrepôt de données orienté XML. Les documents sont stockés au format XML afin de profiter de la technologie d'Oracle XML.

D'autres initiatives telles que DW4TR, ERS Kyoto et NC Catch ont préféré modéliser leur entrepôt sur un modèle multidimensionnel afin de pouvoir utiliser les technologies OLAP (OnLine Analytical Processing) (Hu et al., 2011; Studnicki et al., 2010; Yamamoto et al., 2012).

Quand à Emerse, il est basé sur un entrepôt de document indexé par Lucene (Hanauer, 2006).

Des travaux spécifiques au domaine médical ont défini des modèles permettant d'améliorer l'interopérabilité entre les systèmes. L'initiative openEHR (Kalra et al., 2005) propose de séparer le modèle de référence du modèle physique. Ils ont défini la notion d'archétype pour modéliser chaque donnée atomique du dossier médical et le langage d'interrogation correspondant AQL (Archetype Querying Language). Ils mettent un ensemble de librairies et d'outils en open source pour la communauté. Archetype-based DWH (Marco-Ruiz et al., 2015) a repris le principe d'openEHR pour développer son modèle de données.

SMEYEDAT est développé sur un modèle relationnel centré sur le patient, les faits sont dispersés dans des tables spécifiques à chaque catégorie de faits (Diagnostics, procédures etc.). Il est peu évolutif, mais très spécifique à l'ophtalmologie, discipline pour laquelle il a été développé (Kortüm et al., 2017).

2.3. Les types de données intégrés

Les 16 entrepôts de données analysées utilisent des données déidentifiées, voire anonymisées (CATCH-IT). Certains entrepôts permettent de ré-identifier les patients afin de les inclure dans des études cliniques (i2b2, Stride, Vanderbilt, Radbank). D'autres sont totalement nominatifs car utilisés au quotidien par les cliniciens qui prennent en charge ces patients (SMEYEDAT, Radbank). Il s'agit dans les deux cas d'entrepôts de données spécifiques à une discipline médicale (ophtalmologie et radiologie).

Excepté Emerse qui est davantage un « document repository », les entrepôts étudiés sont alimentés avec des données codées issues des résultats biologiques, de la facturation (diagnostic CIM10), de formulaires cliniques spécifiques ou de formulaires recherches (DW4TR). Stride et Mayo Clinic intègrent les données issues du DPI, utilisant le modèle HL7 RIM. L'intégration de ces données nécessite des thesaurus spécifiques permettant de coder les données dans l'entrepôt

de données structurées. Malgré une volonté de s'aligner sur des référentiels nationaux ou internationaux, certains entrepôts de données (DW4TR) sont restés sur des thesaurus locaux leur permettant d'intégrer rapidement les données sans passer par une étape d'alignement terminologique. Les auteurs de DW4TR prévoient toutefois de passer à terme à un référentiel national commun.

Plus de la moitié des entrepôts de données étudiés disent intégrer les données en texte libre. Mayo Clinic, Consore et Meteor n'intègrent pas directement le texte libre mais les concepts extraits.

Les données -omiques sont généralement intégrées dans un environnement séparé, avec la possibilité de connecter les informations en fonction des politiques d'accès mises en œuvre (Brammen et al., 2005; Roden et al., 2008).

2.4. Les fonctionnalités des entrepôts de données

La recherche de patients

Les interfaces mises en œuvre pour retrouver les patients dépendent du type de données intégrées, des utilisateurs (médecin, chercheur, data scientist) et du degré d'intégration de l'entrepôt dans une interface Homme Machine.

Les méthodes développées dans la discipline de « l'information retrieval » ne s'appliquent pas aux moteurs de recherche développés dans les entrepôts de données biomédicaux étudiés. Il n'y a pas de notion de distance ou de ranking, lié à un calcul de similarité entre la requête de l'utilisateur et les documents recherchés. Les critères de recherche sont appliqués de manière booléenne et les patients retrouvés doivent répondre à tous les critères spécifiés.

i2b2 propose un moteur de recherche basé sur les données codées. L'interface permet une navigation au sein des thesaurus utilisés pour décrire les faits. La construction de la requête se faire par un glisser déposer dans les colonnes définies comme des groupes de requête (une manière ingénieuse de gérer les ET et les OU entre les groupes de requêtes). Pour certaines données codées la valeur que peut prendre le concept sélectionné peut être précisée (intervalle, borne, liste de valeurs possibles). Des critères supplémentaires peuvent être ajoutés à chaque groupe de requêtes (intervalle de temps, critère d'exclusion, contraintes temporelles entre des évènements associés aux groupes). Plusieurs types de résultats sont disponibles : le nombre de patients, la liste des patients, les timeline des patients etc. Tous les patients retrouvés sont considérés comme vrais positifs. Jannot *et al.* ont décrit l'utilisation de l'entrepôt de données i2b2 de l'Hôpital Européen Georges Pompidou sur 8 années (Jannot et al., 2017). Malgré un succès certains lié aux compétences de l'équipe d'informatique médicale, les utilisateurs ont des difficultés à manipuler l'interface. Les auteurs préconisent l'accompagnement d'un bio-informaticien pour la constitution des cohortes.

Excepté Emerse, tous les entrepôts de données étudiés présentent ces fonctionnalités de navigation dans un ou plusieurs thesaurus afin de sélectionner les concepts permettant de rechercher les patients.

i2b2, Radbank, Stride, Smeyedat, Vanderbilt et Emerse permettent de rechercher directement dans le texte libre. Seul Emerse insiste sur l'importance de proposer une interface permettant de visualiser les données qui ont permis de retrouver les patients (par exemple, en affichant l'extrait du texte correspondant). Il permet de vérifier la véracité de chaque cas retrouvé car la recherche en texte libre peut renvoyer des patients faux positifs. Le moteur de recherche utilise la technologie Lucene pour indexer les documents ("Apache Lucene - Apache Lucene Core," n.d.). Les utilisateurs définissent les listes de termes qu'ils veulent rechercher en précisant les synonymes et éventuellement les expressions exclusives (par exemple des tournures de phrases négative : « absence de diabete »). Les groupes de termes (« bundle ») peuvent être sauvés et partagés ce qui permet de mutualiser les requêtes complexes. Cela n'en reste pas moins laborieux de devoir définir à chaque recherche les critères d'exclusion permettant de limiter le bruit lié à la négation ou aux antécédents familiaux. L'expansion terminologique est basée sur le même principe, en listant dans le groupe de termes tous les synonymes et déclinaisons possibles.

Stride et Radbank utilisent le module Oracle Text disponible dans le SGBD Relationnelle Oracle. La négation n'est pas prise en compte par Radbank. Les auteurs de Stride ne précisent pas si la recherche en texte libre prend en compte la négation ou les antécédents familiaux du patient.

i2b2 et Vanderbilt utilisent des requêtes SQL classiques pour interroger le texte libre (*like*). Smeyedat utilise une solution industrielle Qlickview. Escudié *et al.* ont développé un module complémentaire à i2b2 permettant de rechercher dans le texte libre et de visualiser si les cas retrouvés sont vrais positifs (Escudié et al., 2015).

Autres fonctionnalités

Certains entrepôts de données proposent des fonctionnalités supplémentaires au moteur de recherche. i2b2 permet de visualiser les statistiques démographiques des patients retrouvés dans les résultats d'une recherche. Il permet aussi de comparer deux populations issues de deux requêtes.

Smeyedat est spécialisé en ophtalmologie, il propose dans son interface des indicateurs spécifiques à cette discipline. Consore permet de créer des paniers (cohorte) de patients, de partager des requêtes, d'interroger plusieurs sites simultanément (équivalent de SHRINE d'i2b2) (Livartowski, 2016). Meteor intègre un modèle prédictif calculant le risque de réadmission des patients en se basant sur des méthodes de Vector Support Machine. DW4TR, NC CATCH et ERS proposent une interface de type OLAP calculant des indicateurs à

l'intersection des dimensions présentes dans le modèle de données. L'entrepôt de Vanderbilt permet d'alimenter automatiquement une instance RedCap (Harris et al., 2009), évitant la ressaisie de données.

La similarité entre patients

Aucun des entrepôts de données analysés précédemment n'intègre la recherche de patients similaires. La plupart des travaux menés sur les calculs de similarité entre patients ont été faits à partir des données codées à la source.

Montani *et al.* utilisent les données mesurées pendant la dialyse des patients afin de retrouver des sessions de dialyse similaires à un patient index. Leur méthode a utilisé des modèles mathématiques permettant d'intégrer des séries temporelles dans leur calcul de distance (Montani et al., 2006). Sun *et al.* intègrent dans la mesure de similarité basée sur la distance de Mahalanobis les retours des médecins améliorant ainsi leur algorithme au fur et à mesure de son utilisation (Sun et al., 2012). Vallati *et al.* ont développé une méthode de similarité basée sur la distance euclidienne en utilisant 4 variables cliniques codées (Vallati et al., 2013).

PhenomCentral (Buske et al., 2015) est un portail dédié aux maladies rares permettant aux patients et aux médecins de décrire leurs signes cliniques avec le thesaurus Human Phenotype Ontology (HPO) (Groza et al., 2015). Ils ont intégré des scores de similarité entre patients basés sur la distance sémantique des concepts HPO. Un des scores évalué est PhenoDigm (Smedley et al., 2013). Il intègre les distances sémantiques entre tous les concepts phénotypiques de deux entités (patient, syndrome, modèle animal, gène etc.). La distance sémantique est calculée avec le script OWLSim introduit pour la première fois par Chen *et al.* (Chen et al., 2012) (http://owlsim.org).

GeneYenta est un outil de case matching utilisant les algorithmes développés pour les sites de rencontre (Gottlieb et al., 2015). Les auteurs insistent sur l'amélioration de la méthode par la pondération des cliniciens sur les « matchs » proposés par le système.

Lee *et al.* ont développé une méthode de prédiction de risque de mortalité pour un patient à partir des patients similaires, la distance de similarité est basée sur le Vector Space Model, en intégrant les données structurées de MIMIC II (biologie, ICD9) (Lee et al., 2015).

2.5. Les méthodes de traitement automatique du langage dans l'intégration des données textuelles

La reconnaissance de concepts

L'extraction phénotypique ou la reconnaissance de concepts dans les comptes rendus médicaux a été initiée avec le Linguistic String Project-Medical Language Processor (LSP-MLP) (Sager et al., 1986), suivi par le projet MEDLee (Medical Language Extraction and Encoding system) (Barrows Jr et al., 2000; Friedman, 1997; Friedman et al., 2004), et MetaMap (Aronson, 2001). Ces

outils open source utilisent le Metathesaurus® du Unified Medical Language System® (UMLS) comme dictionnaire afin de reconnaitre la liste la plus exhaustive des concepts présents dans un texte médical. L'UMLS est développé par le National Library of Medicine américaine (Lindberg et al., 1993). Il propose plusieurs ressources terminologiques : le SPECIALIST Lexicon pour le traitement automatique du langage ("Fact SheetSPECIALIST Lexicon," n.d.), le Semantic Network pour associer les concepts à un réseau de types sémantiques reliés, par relations hiérarchiques et sémantiques ("Fact SheetUMLS® Semantic Network," n.d.), et l'UMLS Metathesaurus contenant l'ensemble des termes médicaux ("Fact SheetUMLS® Metathesaurus®," n.d.). L'UMLS développe aussi des outils utilisant leurs ressources dont MetaMap.

D'autres initiatives ont utilisé l'UMLS pour extraire des concepts. Long propose une méthode à base de règles permettant de retrouver les diagnostics et les procédures en utilisant les termes à proximité (verbes, mots, ponctuation) (Long, 2005). Bashyam *et al.* ont développé un module de MetaMap simplifiant la tokenisation et réduisant largement le temps de process. Ils précisent que leurs résultats reflètent la qualité des données du thesaurus utilisé, et qu'un certain nombre de faux positifs sont dus à des erreurs dans l'UMLS (Bashyam et al., 2007). Wu *et al* ont extrait les concepts UMLS sur 51 millions de comptes rendus cliniques de Mayo Clinic pour évaluer la distribution lexicale dans leur corpus et ainsi créer un thesaurus adapté à leur usage local (Wu et al., 2012).

cTAKES (Savova et al., 2010) utilise le Stanford Named Entity Recognizer (NER) (Finkel et al., 2005), un algorithme basé sur des méthodes de machine learning.

Divita *et al.* ont développé une alternative à MetaMap et cTakes qui est 18 fois plus rapide que cTakes et 7 fois plus que MetaMap en limitant le prétraitement du texte (Divita et al., 2014). Toutefois la précision en est diminuée.

Meystre *et al.* décrivent avec précision toutes les étapes nécessaires à l'extraction d'information depuis un compte rendu clinique (Meystre et al., 2008). Ils insistent sur la complexité de cette tâche en raison de l'hétérogénéité de ce média, hétérogénéité de longueur, de qualité grammaticale, de qualité orthographique, du vocabulaire spécifique (acronyme et abréviation) etc. Une partie importante de la littérature du domaine porte sur le prétraitement du texte avant l'extraction d'information : la correction des fautes d'orthographe (particulièrement présente dans les textes cliniques) (Ruch et al., 2003), le découpage en phrase et la tokenisation du texte (Tomanek et al., 2007), l'étiquetage morphosyntaxique (part-of-speech tagging) (Campbell and Johnson, 2002, 2001), la désambigüisation du texte (notamment pour les acronymes (Pakhomov et al., 2005)).

Wei et Denny positionnent l'extraction d'information dans la recherche génétique et sur la nécessité de réutiliser le dossier de soins pour améliorer la description phénotypique des

mutations génétiques (Wei and Denny, 2015). Ils insistent sur la difficulté à obtenir un phénotypage de bonne qualité, reprenant les arguments de Meystre précédemment cité.
Ces outils de reconnaissance de concepts ou d'extraction de concepts se sont enrichis pour associer la possibilité de classer le concept suivant le *contexte* (histoire du patient, antécédents familiaux, etc.) dans lequel il est utilisé ainsi que le niveau de *certitude*.

La détection de la négation

La plupart des méthodes de traitement automatique du langage publiées concernent l'extraction et la classification des concepts médicaux. Dés 2001, plusieurs travaux ont utilisé les méthodes basées sur les expressions régulières [8,9]. L'outil le plus cité est Negex développé pour l'anglais par Chapman en 2001 (Chapman et al., 2001a). La méthode consiste à classer en « négatif » ou « non négatif » un concept médical présent dans une phrase en fonction de la présence ou non d'une expression de la négation avant ou après ce concept. A partir d'un corpus d'apprentissage constitué de courriers de sorties, Chapman décrit 35 manières d'exprimer la négation sous forme d'expressions régulières. Elle considère 2 groupes d'expression de la négation : les « pseudo négations » (10 expressions) qui correspondent aux doubles négations et ne doivent donc pas être prises en compte, les « vraies négations » (25 expressions) qui sont utilisées pour nier un signe ou une maladie. Les expressions régulières de type « vraie négation » sont elles même séparées en 2 sous groupes correspondant à la position de la négation avant ou après les concepts niés et éloignés de maximum 5 mots. Elle évalue son système en classant une liste de concepts médicaux issus du UMLS, dont les types sémantiques UMLS sont "Finding", "Disease or Syndrome", ou "Mental or Behavioral Dysfunction". Sur 1000 phrases, NegEx classe 1235 concepts UMLS avec une sensibilité de 78% et une spécificité de 95%. Malgré les améliorations possibles, ce travail montre l'efficacité des expressions régulières pour détecter simplement l'expression de la négation dans des textes. D'autres publications ont par la suite confirmé l'intérêt de cette approche (Goryachev et al., 2006; Huang and Lowe, 2007; King et al., 2011). NegEx a par la suite été enrichi pour des problématiques ciblées comme l'analyse des syndromes grippaux dans les courriers cliniques des vétérans américains (South et al., 2007). Des équipes de recherche l'ont adapté à d'autres langues notamment le suédois et le français (Chapman et al., 2013; Skeppstedt, 2011). Parallèlement à cela, NegEx a été intégré dans différents projets plus généraux sur l'analyse sémantique des comptes rendus : ConText en 2009 (Harkema et al., 2009), cTakes en 2010 (Savova et al., 2010), DEEPEN en 2014 (Mehrabi et al., 2015). Ces derniers travaux ont ajouté à la négation de nouvelles classes : l'hypothétique, l'historique (antécédent) et non présent chez le patient (antécédents familiaux). D'autres méthodes combinant des expressions régulières et le Machine Learning sont publiées à partir de 2011 (Clark et al., 2011; Wu et al., 2014). Wu conclut que pour obtenir des résultats pertinents ces méthodes nécessitent des corpus annotés

d'apprentissage spécifiques à chaque domaine d'application (radiologie, cardiologie etc.) et à chaque langue.

La détection des antécédents familiaux

Comme pour la négation, la plupart des travaux publiés sont des méthodes de classification de concepts médicaux comme étant reliés au patient ou non reliés au patient. Les premiers travaux se sont basés sur la structuration du compte rendu clinique en section. Friedlin & McDonald (Friedlin and McDonald, 2006) se sont concentrés sur la détection d'une section identifié par un titre indiquant la notion d'antécédents familiaux. Ils ont considéré que les phrases de cette section concernés les antécédents familiaux du patient. Ils ont utilisé REX (Regenstrief data eXtraction tool) pour classer 12 maladies dans ces phrases avec une précision et un rappel de 97% et 96%. Goryachev *et al.* sont aussi partis du principe que les comptes rendus hospitaliers sont structurés en sections, et que les sections sont décrites par un titre permettant de retrouver la section « antécédents familiaux » (Goryachev et al., 2006). Ils ont utilisé l'UMLS pour extraire les concepts médicaux et les classés par type sémantique. Les auteurs ont développé un set de règles pour associer les concepts médicaux aux membres de la famille du patient ou au patient. Leur évaluation donne une sensibilité de 97,2% et une spécificité de 99,7%. Lewis et al. (Lewis et al., 2011) prennent en compte l'intégralité du compte rendu, considérant que les antécédents familiaux peuvent être disséminés dans le texte. Ils utilisent le parser de Stanford pour détecter les dépendances entre les maladies et les liens familiaux avec une précision de 61% et un rappel de 51%.

L'algorithme ConText développé par Harkeman et al. (Harkema et al., 2009) est basé sur l'approche NegEx. Il est basé sur des règles à base d'expressions régulières ; De la même manière que pour la classification de la négation, il va déterminer si un concept médical retrouvé dans une phrase se rapporte à une négation, une hypothèse, aux antécédents du patient ou à un autre individu. Les auteurs ont montré encore une fois la bonne performance des expressions régulières pour classer des concepts comme étant des antécédents familiaux ou non.

2.6. Les objectifs

Notre objectif est de relever les défis identifiés par Prokosch pour la réutilisation des données de soin (Prokosch and Ganslandt, 2009) :

- Établir des entrepôts de données cliniques complets en exploitant et en intégrant les données cliniques des patients à partir d'une multitude de bases de données cliniques.
- Développer des méthodologies d'extraction innovantes afin d'obtenir de nouvelles informations et connaissances.
- Créer un environnement supportant la recherche translationnelle

Nous proposons de développer un entrepôt de données répondant aux critères de Prokosch et permettant deux niveaux d'intégration :

(1) un niveau à coût réduit permettant d'intégrer rapidement les sources quelque soient leur format.

(2) un niveau à coût plus élevé d'intégration des données structurées permettant de requêter des données fiabilisées

Nous proposons également une interface homme machine proche du clinicien, lui permettant une autonomie maximale pour recruter des patients, explorer les données phénotypiques, diagnostiquer de patients, et un modèle de données adapté à l'intégration de données textuelles et structurées.

3. Méthodes

3.1. Un modèle d'entrepôt de données orienté document

Comme nous l'avons déjà évoqué en introduction, le texte libre est l'élément central du DPI. Mais ce n'est pas le seul argument en faveur d'un entrepôt orienté document. Wisniewski a souligné la difficulté d'intégrer plusieurs sources de données dans un entrepôt de données structurées car cela nécessite de connaître la structure des données sources, d'en comprendre leur particularité et leur subtilité (Wisniewski et al., 2003). Un temps considérable est nécessaire pour réaliser l'alignement terminologique entre les sources. Et le résultat est rarement satisfaisant. Le traitement automatique du langage permet d'accélérer cet alignement, cela n'en reste pas moins une tâche couteuse (Vuokko et al., 2017). Brandt et al. (Brandt et al., 2002) puis Dinu en 2007 (Dinu and Nadkarni, 2007) ont montré la difficulté que représente le modèle en étoile pour la représentation de relations complexes entre les données.
Notre objectif est d'intégrer toutes les données d'un hôpital, quelque soit le logiciel qui a produit cette donnée et quelque soit la période à laquelle il l'a produite. Orienter l'entrepôt autour de la notion de document nous permet d'être totalement interopérable avec toutes les sources de données. Certaines sources de données sont déjà codées avec des thesaurus formels (CIM10) ou locaux (résultats biologiques), il s'agit de conserver ces informations structurées dans l'entrepôt de données.

Nous définissons un document comme un texte produit pour un patient à une date spécifique et par une entité hospitalière (service, unité). Il s'agit des lettres et des comptes rendus, les résultats des examens complémentaires (biologie, radiologie, imagerie), des prescriptions (*Code de la santé publique - Article R1112-2*, n.d.).
Certaines données médicales n'existent pas sous forme de compte rendu dans le dossier patient, mais uniquement sous forme de données codées. Par exemple, les codes diagnostics CIM10 utilisés pour la Tarification à l'activité ou les actes médicaux codés en CCAM (Classification Commune des Actes médicaux) n'existent que sous la forme de code. Pour que ces données puissent être intégrées dans notre entrepôt, nous créons un document ad-hoc formé de la concaténation des codes, des libellés et éventuellement des valeurs associées (par exemple « principal », « associé » pour les diagnostics). Le document doit regrouper des codes qui correspondent à la même date, produits par la même structure hospitalière. Par exemple, pour la CIM10, nous créons un document concaténant les diagnostics par Résumé d'Unité Médicale (RUM). Le double intérêt est de pouvoir d'une part afficher un document facile à lire et d'autre part le retrouver aussi bien à partir d'une requête structurée que d'une requête plein texte.

3.1.1. Pourquoi créer un nouveau modèle ?

Afin de remettre le compte rendu médical au centre de l'entrepôt de données, nous avons décidé de partir d'un nouveau modèle de base de données. Nous aurions pu repartir du modèle i2b2 considéré comme un standard aujourd'hui. Mais ce choix ne nous a pas paru satisfaisant. Nous voulons proposer un système proche du clinicien, de sa manière de travailler et de sa manière de penser. Cela constitue un changement de paradigme et de domaine d'application par rapport à i2b2. Cela implique un stockage des données adapté et optimisé. Par exemple, i2b2 ne permet pas aujourd'hui de regrouper des faits autour d'un document de manière simple et naturelle. Ce serait toutefois possible en utilisant la notion de *modifier* et de numéro d'instance. Même s'il conserve une interopérabilité sur des fonctionnalités simples en supposant un référentiel commun, l'ajout des *modifiers* limite l'interopérabilité quant au stockage des données et à leur interprétation.
Le gain obtenu par la conservation du modèle i2b2 était en deçà du gain de lisibilité et d'optimisation du modèle que nous proposons.

3.1.2. Le modèle de données

L'annexe 1 décrit le schéma relationnel de manière simplifiée.
Le modèle est constitué de 4 types de tables :

- les tables contenant les données administratives des patients
- les tables contenant les données médicales des patients
- les tables contenant les thesaurus
- les tables dédiées au fonctionnement de l'entrepôt (données des utilisateurs, tables temporaires). Nous ne décrirons pas ici en détail ces tables.

La table contenant les documents est situé au centre du modèle.

Les tables décrivant les patients:
La table DWH_PATIENT contient les informations administratives du patient (identité, adresse, lieu de naissance). Les données géographiques sont aussi stockées au format latitude et longitude en coordonnées Lambert. L'entrepôt de données étant à destination des cliniciens, il est indispensable de conserver les données nominatives dans l'entrepôt.

La table DWH_PATIENT_IPPHIST permet de conserver l'historique des identifiants que prend un patient d'une source de données à une autre. Par ailleurs, au sein d'une même source de données, le patient peut avoir plusieurs identifiants parce qu'il a été créé en double puis fusionné. Une colonne permet de préciser lequel doit être affiché pour l'utilisateur.

La table DWH_PATIENT_REL permet d'indiquer des liens de parentés entre deux patients de l'entrepôt. Cette table est particulièrement utile dans le cas d'un hôpital comme Necker

spécialisé dans les maladies rares pour lesquelles les études familiales sont indispensables pour diagnostiquer le patient.
Les sources de données non institutionnelles ne contiennent pas l'identifiant hospitalier du patient, notamment les sources de données anténatales (échographie, foetopathologie) pour lesquelles les fœtus sont rattachés à la mère et n'ont pas d'existence administrative. Il a fallu intégrer dans le processus ETL (Extract, Tranform and Load) une méthode de séparation des données de la mère et du foetus, et une réconciliation des données du fœtus et de l'enfant, s'il a été vu à l'hôpital après sa naissance. Nous avons mis en place une stratégie de réconciliation des identités des patients par des méthodes déterministes à base de règles en utilisant la base d'identité institutionnelle (Dusetzina et al., 2014; Grannis et al., 2002).

Les tables contenant les données médicales:
La table DWH_DOCUMENT contenant les documents est au centre du schéma en étoile, chaque document est identifié par une clé unique. Des métadonnées sont directement intégrées dans cette table et permettent de décrire le document de manière succincte : la date du document, l'auteur, le service producteur, la clé du document dans le logiciel source, le logiciel source. Ces métadonnées sont répétées dans les tables DWH_DATA et DWH_TEXT décrites plus bas. La redondance de ces informations permet d'améliorer les performances du moteur de recherche en réduisant les jointures SQL (Chen, 2001). La volumétrie supplémentaire nécessaire est négligeable et peu coûteuse aujourd'hui. Le document inséré dans cette table correspond au document qui sera affiché à l'utilisateur. Le texte doit être converti au format ASCII avant son insertion dans la table.

La table DWH_TEXT est dédiée à la recherche en texte libre sur les documents. Elle contient le document prétraité pour améliorer la qualité du moteur de recherche. Cette table permet de stocker le document découpé en sous documents correspondant à des contextes et des niveaux de certitude. Le processus est détaillé dans la section 3.2.2. La colonne contenant le texte libre et la colonne contenant le texte libre enrichi sont indexées par Oracle Text pour la recherche textuelle. Les métadonnées de la table DWH_DOCUMENT sont répétées dans cette table (âge du patient, date du document, unité médicale etc.), afin de réduire les jointures lors des requêtes.
Oracle Text permet d'indexer des colonnes pour réaliser des requêtes en texte libre sur ces colonnes. La commande SQL recherchant le texte libre renverra tous les enregistrements de la table correspondant à l'expression recherchée. Cela signifie que si l'utilisateur recherche « expression A AND expression B », les expression A et B doivent se trouver dans le même enregistrement de la table. Donc, si un texte est découpé en phrases et que chaque phrase est stockée dans un enregistrement séparé de la table, la requête ne renverra pas le document dans

lequel les expressions A et B sont séparées dans deux phrases. C'est pourquoi dans la phase ETL nous regrouperons dans un même enregistrement les syntagmes ou phrases répondant à un même contexte et un même niveau de certitude. En fonction des contextes que l'on souhaite créer, une phrase peut être présente dans plusieurs enregistrements de la table DWH_TEXT. Les textes contenus dans cette table sont anonymisés afin d'éviter de donner la possibilité de rechercher des documents par le nom de patients. Particulièrement dans le cas où l'utilisateur n'a pas le droit de visualiser les données détaillées. L'anonymisation se fait par expression régulière à partir des données administratives du patient (nom, prénom, téléphone, date de naissance, adresse).

La table DWH_DATA, équivalente à la table OBSERVATION_FACT d'i2b2 (Klann et al., 2016), contient les données structurées (résultat de laboratoire, PMSI etc.). Chaque donnée structurée est liée à un document dans la table DWH_DOCUMENT. Les données structurées sont codées par des concepts contenus dans la table thesaurus DWH_THESAURUS_DATA (l'équivalent de la table DIMENSION d'i2b2). Les thesaurus peuvent être LOINC, CIM10 ou un thesaurus local à l'établissement.
Certaines sources de données produisent des données codées qui ne sont pas associées à un document ou compte rendu. Par exemple, les Résumés d'Unité Médicale (RUM) sont une liste de codes diagnostics issus du thesaurus CIM10 associés à un type de diagnostic (principal, relié, etc.). Pour respecter le paradigme initial, un document ad hoc est créé dans le processus ETL de ces données. Il est formé de la concaténation des codes et libellés CIM10 produits par une unité hospitalière, à une date et pour un patient. Il est inséré dans les tables DWH_DOCUMENT pour l'affichage, dans la table DWH_TEXT pour la recherche en texte libre et les codes sont insérés dans la table DWH_DATA pour la recherche codée.

La table DWH_ENRSEM contient les concepts extraits pendant la phase d'enrichissement phénotypique. Chaque concept est relié au document, au contexte et au niveau de certitude. Les concepts sont reliés à la table DWH_THESAURUS_ENRSEM qui contient le thesaurus utilisé.
DWH_DATA contient les données codées obtenues à partir des sources. DWH_ENRSEM contient les données codées obtenues par extraction automatique depuis le texte libre. Nous avons volontairement séparé les deux tables en raison de la qualité respectivement vérifiée et non vérifiée des données qu'elles contiennent.

Les tables contenant les mouvements des patients sont :
DWH_PATIENT_STAY : pour les séjours hospitaliers, c'est à dire l'entrée et sortie de l'hôpital avec le mode d'entrée et de sortie

DWH_PATIENT_MVT : les mouvements intra séjour d'une unité à l'autre
DWH_PATIENT_CONSULTATION : les consultations des patients avec l'unité et la date.

Les tables de type thesaurus :
La table DWH_THESAURUS_DATA contient les thesaurus utilisés pour les données structurées (CIM10, Examens de laboratoire – LOINC, thesaurus local etc.).
La table DWH_THESAURUS_ENRSEM contient les thesaurus utilisés pour l'enrichissement terminologique.

Les tables de fonctionnement de l'entrepôt :
Les autres tables concernent le fonctionnement de l'entrepôt de données indépendamment des données qu'il contient (historique des requêtes, gestion des utilisateurs et droits d'accès, gestion des cohortes, journal des accès etc.).

3.2. La recherche d'information

Weng *et al* ont montré le fort potentiel des entrepôts de données cliniques comparé aux registres cliniques pour le recrutement de patients dans les essais cliniques, avec un gain de temps de 20% et une augmentation du nombre de recrutements de 50% (Weng et al., 2010). Un des premiers cas d'usage des entrepôts de données cliniques est la capacité à retrouver des patients par un moteur de recherche (Shin et al., 2014).
La plupart des moteurs de recherche sur les entrepôts de données biomédicaux se font sur les données codées (Résultats de laboratoire, codage PMSI, les données structurées du dossier patient etc.) en raison de leur standardisation. La liste des patients retrouvés peut être directement utilisée pour un projet de recherche sans contrôle manuel.
Cependant, des études récentes ont montré qu'une importante partie des informations restent dans le texte libre (Raghavan et al., 2014). Malgré l'effort demandé aux cliniciens de coder l'information médicale dans le DPI afin de rendre immédiatement exploitables les données issues du soin, ces cliniciens préféreront écrire du texte libre. Nous pensons qu'il y a trois raisons principales à cela : (1) rédiger l'histoire du patient permet de préciser ses doutes et son raisonnement sémiologique au travers un texte littéraire. Cela a pour conséquence que les données codées sont insuffisantes (Cuggia et al., 2010; Escudié et al., 2015) en particulier dans le cas des maladies rares (Garcelon et al., 2014). (2) De plus, il est souvent plus long et fastidieux de retrouver la case à cocher qui contient l'information que de le saisir directement dans un champ de type texte libre. Le temps passé avec le patient est précieux (Toll, 2012). (3) Et finalement, nous ne codons que ce que nous connaissons, le texte libre permet de saisir les détails qui paraissent non pertinents aujourd'hui mais qui peuvent s'avérer déterminants pour décrire l'histoire naturelle d'une maladie. Limiter la description du patient à des éléments codés

est une perte de données importante pour découvrir de nouvelles connaissances en particulier dans le cas des maladies rares non diagnostiquées.

Nous distinguons deux approches pour rechercher des patients à partir du texte libre dans un entrepôt de données.

L'approche codage a posteriori : il s'agit d'extraire des concepts médicaux à partir d'un dictionnaire puis de classer ces concepts suivant leur contexte (antécédents familiaux ou histoire du patient) et leur degré de certitude (négation, affirmation, suspicion etc.) (Harkema et al., 2009; Savova et al., 2010; South et al., 2007; Xu et al., 2010). Puis de proposer un moteur de recherche basé sur le thesaurus utilisé pour l'extraction de concepts. i2b2 utilise ces méthodes de Traitement Automatique du Langage pour intégrer le texte libre dans le modèle et dans le moteur de recherche (Murphy et al., 2006). Il y a deux désavantages à cette méthode : (1) l'utilisateur perd le lien avec le texte à l'origine de la donnée codée, (2) si un terme n'est pas dans le dictionnaire il ne sera jamais extrait et donc le patient sera impossible à retrouver. Park *et al.* rappellent la difficulté pour les outils tels que MetaMap de traiter les fautes d'orthographes ou les différents types de vocabulaire (médecin versus patient) (Park et al., 2015).

Par exemple, dans l'entrepôt de données de Necker, « maladie de Castleman » renvoie 17 patients, alors que le terme seul « Castleman » renvoie 23 patients. 6 patients n'ont pas « maladie de Castleman » mais une « pathologie de Castleman » ou un « syndrome de Castleman ». Alors que l'UMLS 2015 AB ne contient que « maladie de Castleman » (CUI C0017531). De plus quand on recherche « maladie de castelman » (avec une faute d'orthographe, inversion du E et du L) on retrouve 39 patients et 49 avec uniquement « castelman ». Si nous avions utilisé l'extraction de concepts à partir de l'UMLS nous n'aurions retrouvé que 17 patients sur les potentiels 49.

L'approche recherche plein texte : il s'agit de rester au plus proche du texte d'origine et de permettre d'interroger de manière spécifique le texte libre, en prenant compte les nuances du langage. L'interface doit permettre de visualiser les patients trouvés et de comprendre pourquoi ils ont été trouvés. Nous avons choisi l'approche recherche plein texte qui nous paraît la plus adaptée au cas des maladies rares.

Utiliser un moteur de recherche sur les données en texte libre nécessite plusieurs développements spécifiques décrits dans la littérature (Erinjeri et al., 2009; Hanauer et al., 2015; Rosenbloom et al., 2011) :

- un algorithme de traitement automatique du langage pour prendre en compte la négation et les antécédents familiaux dans la recherche plein texte en français

- un algorithme d'expansion sémantique de la recherche afin de prendre en compte les synonymes et les hyperonymes.
- une interface de contrôle et de validation du résultat

3.2.1. Les critères de recherche

L'objectif du moteur de recherche est de donner la possibilité de retrouver des patients répondant à des critères de recherche. Les critères de recherche peuvent être du texte libre et/ou des données codées/structurées qui peuvent être présents dans un ou plusieurs documents disjoints pour le même patient.

Le moteur de recherche que nous proposons est construit autour de requêtes atomiques. Le nombre de requêtes atomiques n'est pas limité. Chaque requête atomique renvoie les documents correspondant aux critères de recherche. Le moteur de recherche calcule l'intersection des patients retrouvés par chaque les requêtes atomiques. C'est à dire qu'il retrouve les patients dont au moins un compte rendu est retrouvé par chaque requête atomique, il peut s'agir de comptes rendus différents d'une requête atomique à l'autre. Nous avons défini 2 types de requête atomique : texte libre, codée. Afin de prendre en compte le maximum de cas d'usage rencontrés par les utilisateurs nous avons ajouté des filtres démographiques et temporels.

La recherche d'information sur le texte libre

Pour la recherche textuelle, nous nous sommes basés sur le module Oracle Text de Oracle® 11g (compatible oracle 12). Il permet de créer un index sur une colonne spécifique d'une table offrant une recherche en texte libre sur le texte contenu dans cette colonne. Les requêtes sont construites avec les opérateurs booléens (AND, OR, NOT). Plusieurs fonctions permettent de préciser la recherche, par exemple la fonction NEAR précise la distance maximum entre deux mots.

Oracle Text permet de spécifier un « lexer » pour la tokenisation des textes indexés. Il définit le langage utilisé, la sensibilité à la casse, la prise en compte des nombres etc. Nous avons défini le moteur insensible à la casse et aux accents.

Nous avons modifié la liste des stop words par défaut en supprimant les mots utilisés pour des indications cliniques : « non », « sans ». En effet, le mot « non » est très souvent utilisé pour préciser un diagnostic, par exemple « diabète non insulinodépendant », « insuffisance rénale chronique non dialysée ».

Par défaut, la recherche en texte libre est réalisée sur le texte non enrichi, avec une certitude à 1 et le contexte « patient_text » (le contexte par défaut est défini dans le paramétrage de l'application). L'utilisateur peut alors modifier ces paramètres et rechercher dans le texte enrichi, ne pas prendre en compte le niveau de certitude ou changer le contexte.

La recherche en texte libre est réalisée sur les colonnes indexées TEXT et ENRICH_TEXT de la table DWH_TEXT. L'alimentation de la colonne TEXT est expliquée dans la section 3.2.2. L'alimentation de la colonne ENRICH_TEXT est expliquée dans la section 3.2.3.

La recherche d'information sur données structurées

La sélection de critères codés n'est pas basée sur une arborescence. Nous avons préféré utiliser un moteur de recherche en texte libre sur la description des données codées dans la table thesaurus. Cette description est la concaténation du libellé, de l'unité et éventuellement des informations complémentaires fournies par le thesaurus d'origine. Pour les résultats biologiques, les informations complémentaires sont la technique d'analyse et le milieu du prélèvement (sang, urine etc.). La colonne description est indexée par Oracle Text.

L'objectif est de proposer un outil très simple et très ergonomique pour retrouver un concept dans le thesaurus sélectionné (Diagnostic CIM10, Laboratoire, etc.). Le moteur analyse la requête de l'utilisateur et ajoute automatiquement les « and » et les « % » en fin de chaque mot pour être compatible avec une requête Oracle Text. Par exemple dans le thesaurus des résultats biologiques, la recherche « créat sang » va exécuter la requête « creat% and sang% » et renvoyer la liste des examens possibles au sein de l'arborescence du thesaurus.

L'utilisateur sélectionne le critère codé qu'il souhaite ajouter dans la requête atomique. En fonction du type de critère (liste de valeurs, numérique, binaire), le système adapte les filtres affichés à l'utilisateur.

Options supplémentaires dans les requêtes atomiques

Chaque requête atomique est par défaut un critère d'inclusion. L'utilisateur peut le définir comme critère d'exclusion dans les options supplémentaires. Tous les patients ayant un document retrouvé par ce critère seront alors exclus du résultat.

L'utilisateur a aussi la possibilité de préciser des conditions supplémentaires sur les documents recherchés : un intervalle d'âge du patient à la date du document, un intervalle de date du document, la source du document, le service ayant produit le document, un intervalle de temps minimum entre 2 documents contenant le terme recherché ou la donnée structurée.

Des options spécifiques à la recherche en texte libre sont possibles et déjà évoqué précédemment.

Les contraintes de temporalité

Sun et Alonso ont montré l'intérêt d'intégrer la temporalité dans les outils de data mining (Alonso et al., 2007; Sun et al., 2013a). Plusieurs travaux décrivent l'intégration de ces contraintes dans leur système de recherche d'information (Crowley et al., 2010; Plaisant et al., 2008). Nous avons intégré trois types de relation temporelle entre deux requêtes atomiques :

- simultanés (ayant lieu le même jour),
- sont éloignés d'un nombre minimum ou maximum de jours/mois/années
- sont dans un ordre spécifique avec un minimum ou maximum de jours/mois/années les séparant.

Nous avons ajouté l'option « minimum strict » qui permet de préciser qu'il exclut les patients qui ont 2 documents dont la distance est inférieure au minimum indiqué même s'ils ont des documents dont la distance temporelle est au dessus de ce minimum. Par exemple, si l'utilisateur veut les patients avec « insuffisance rénale » dans un document et 3 ans après au minimum (mais pas avant) « dialyse » dans un document, il utilisera alors l'option « minimum strict » plutôt que minimum.

Les contraintes démographiques ou générales

Si les critères démographiques ne représentent que 2,3% des critères d'inclusion dans les essais cliniques publiés dans clinicaltrial.gov (Ross et al., 2010), ils sont présents dans quasiment toutes les études. Nous avons intégré la possibilité de filtrer les patients sur des critères démographiques. Les critères sont :

- Le sexe du patient
- L'intervalle d'âge du patient aujourd'hui,
- Patient décédé ou vivant
- L'intervalle d'âge du patient au décès
- La période de 1ere venue (intervalle de dates)
- L'intervalle d'un nombre minimum d'années de suivi dans l'entrepôt de données
- Les cohortes desquelles les patients sont exclus du résultat

Un format d'échange pour les critères de recherche

Nous avons définit un format XML pour encapsuler l'intégralité des critères d'une requête. Le format XML est présenté en Annexe 2. Le moteur transforme le fichier XML en requêtes SQL. Ce stockage des requêtes permet de faciliter les échanges, de transférer des requêtes et de les exécuter (Weber et al., 2009). On pourra envisager une interopérabilité multi site en supposant une interopérabilité des thesaurus de données codées. Pour les requêtes atomiques de type texte libre, les requêtes sont directement interopérables sur un autre entrepôt utilisant la même langue.

3.2.2. Prise en compte de la négation et des antécédents familiaux

Les comptes rendus hospitaliers contiennent une description détaillée des signes et symptômes du patient ainsi que les absences de signes ou l'élimination de certains diagnostics ou gènes candidats. C'est pourquoi sans prétraitement, une recherche plein texte peut renvoyer un grand

nombre de résultats faux positifs liés à l'utilisation de la négation (Chapman et al., 2001b; Chu et al., 2006; Wu et al., 2011).

Dans le contexte des maladies rares et des maladies génétiques, les comptes rendus décrivent les informations cliniques des membres de la famille du patient sur plusieurs générations. Par exemple, dans l'entrepôt de données de Necker, la recherche plein texte « crohn » renvoie 2095 patients, alors qu'en réalité seulement 1200 patients ont été identifiés avec une maladie de crohn. Les 800 patients faux positifs retrouvés ont un membre de la famille atteint de la maladie de crohn sans être eux même atteints.

Notre objectif est donc de pouvoir rechercher dans les comptes rendus en texte libre en écartant les éléments du texte évoquant les antécédents familiaux du patient ou considérés comme une négation. Pour cela nous proposons de développer un algorithme classant les phrases d'un texte en deux catégories : patient et antécédents familiaux, puis dans chaque catégorie, de classer les syntagmes en deux catégories : les syntagmes non négatifs et les syntagmes négatifs. Le pipeline est composé de 4 phases :

- Il a fallu développer une stratégie de reconstruction des phrases en utilisant la ponctuation, les majuscules et le nombre de mots après un saut de ligne.
- Pour chaque phrase la présence d'un mot correspondant à l'expression d'un membre de la famille classait la phrase comme dans les antécédents familiaux
- Nous avons défini une liste de termes et de règles pour découper les phrases en syntagmes, en considérant qu'une négation ne valait pas pour toute la phrase
- Pour classer chaque syntagme en négation ou non négation, nous avons utilisé une liste d'expressions régulières exprimant la négation à partir d'un corpus de comptes rendus cliniques (différents de l'évaluation). Les expressions régulières ont été enrichies avec des règles tenant compte des doubles négations et de la normalité (considéré comme négation : NPHP1 normal).

Nous avons évalué notre capacité à réduire le bruit du moteur de recherche en comparant la précision et le rappel avec et sans la prise en compte de la négation, avec et sans la prise en compte des antécédents familiaux, sur trois requêtes : « crohn and diabete », « lupus and diarrhee », « nphp1 ». Nous améliorons la précision de 28% à 85%, le rappel est diminué de 100% à 98% et la F-measure passe de 43% à 91%.

Notre stratégie basée sur les expressions régulières montre une excellente capacité à améliorer le moteur de recherche. Toutefois, nous n'avons pas intégré la capacité à détecter la suspicion ou la « recherche de ». L'article ci dessous détaille notre méthodologie.

Journal of the American Medical Informatics Association, 24(3), 2017, 607–613
doi: 10.1093/jamia/ocw144
Advance Access Publication Date: 20 October 2016
Case Report

Case Report

Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse

Nicolas Garcelon,[1,2] Antoine Neuraz,[1,2] Vincent Benoit,[1] Rémi Salomon,[1,3] and Anita Burgun[2,4]

[1]Institut Imagine, Paris Descartes Université Paris Descartes-Sorbonne Paris Cité, Paris, France, [2]INSERM, Centre de Recherche des Cordeliers, UMR 1138 Equipe 22, Université Paris Descartes, Sorbonne Paris Cité, Paris, France, [3]Service de Néphrologie Pédiatrique, Hôpital Necker-Enfants Malades, Assistance Publique -Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France, and [4]Hôpital Européen Georges Pompidou, Assistance Publique -Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France

Corresponding Author: Nicolas Garcelon, 24 boulevard du Montparnasse, 75015 Paris, France. E-mail: nicolas.garcelon@institutimagine.org; Tel: 0033.1.42.75.44.57

Received 25 January 2016; Revised 12 August 2016; Accepted 31 August 2016

ABSTRACT

Objective: The repurposing of electronic health records (EHRs) can improve clinical and genetic research for rare diseases. However, significant information in rare disease EHRs is embedded in the narrative reports, which contain many negated clinical signs and family medical history. This paper presents a method to detect family history and negation in narrative reports and evaluates its impact on selecting populations from a clinical data warehouse (CDW).

Materials and Methods: We developed a pipeline to process 1.6 million reports from multiple sources. This pipeline is part of the load process of the Necker Hospital CDW.

Results: We identified patients with "Lupus and diarrhea," "Crohn's and diabetes," and "NPHP1" from the CDW. The overall precision, recall, specificity, and F-measure were 0.85, 0.98, 0.93, and 0.91, respectively.

Conclusion: The proposed method generates a highly accurate identification of cases from a CDW of rare disease EHRs.

Key words: data warehouse, search engine, natural language processing, rare diseases, electronic health records

INTRODUCTION

Secondary use of electronic health records (EHRs) for research purposes requires information retrieval tools to make structured data and information contained within clinical text accessible. Narrative clinical reports allow flexibility of expression and representation of elaborate clinical entities, including clinical signs, patient history, and family history.[1–7]

In order to facilitate this needed capacity for data exploration at our institution (Necker Enfants Malades and Imagine Institute), we have designed and deployed Dr Warehouse®, a full-text clinical data warehouse (CDW) for cohort identification and data extraction. Necker/Imagine specializes in rare diseases. A detailed description of the symptoms, including negative findings, is documented in the patient record, as it is required to lead to an accurate diagnosis. Therefore, it is crucial to include mechanisms for detecting negation in text. A narrative rare-disease patient report includes a large amount of detailed information from 3 generations of relatives. Exploiting this information and distinguishing between family history and patient information is crucial not only for monogenic disorders (e.g., NPHP1 mutation, responsible for steroid resistant nephrotic syndrome) but also for all diseases with a known hereditary component

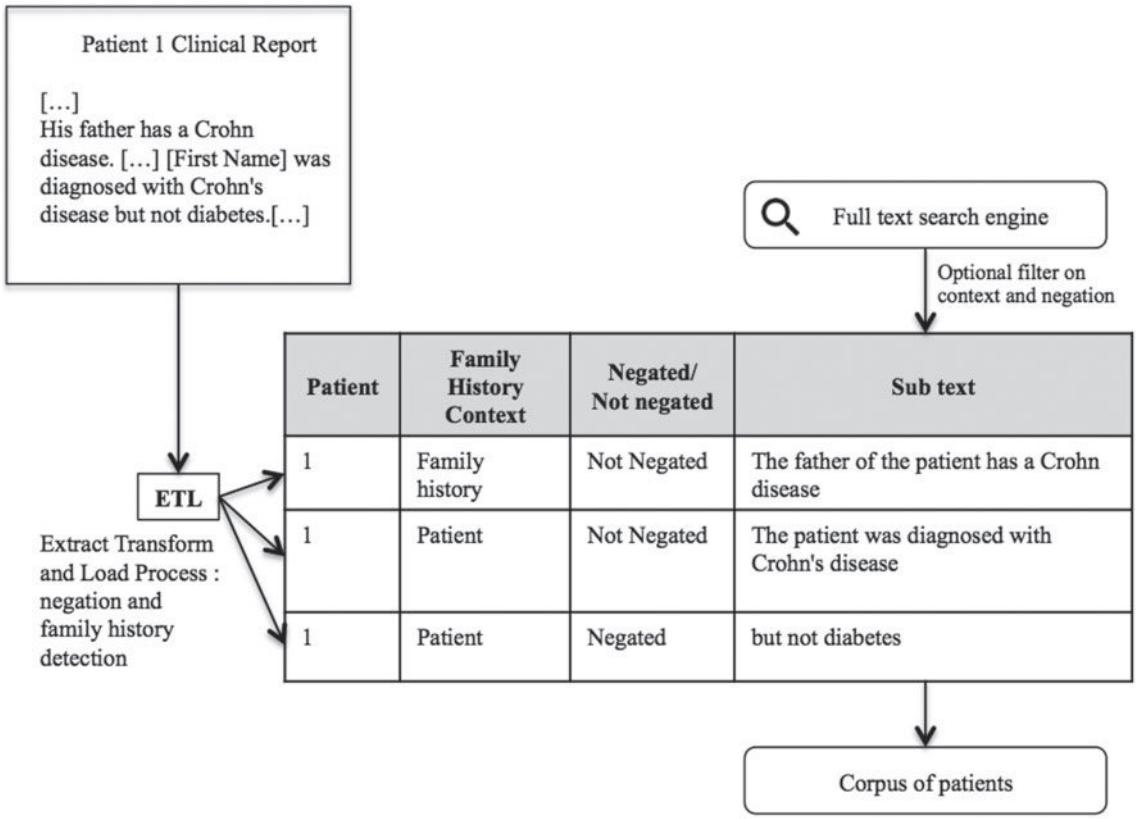


Figure 1. Illustration of the objective of our method and its context.

(e.g., Crohn's disease and lupus).[8,9] Moreover, patients who suffer from 1 autoimmune disease are more likely to present other autoimmune disorders. Genomewide association studies have identified several genes that might be associated with increased susceptibility to diabetes and Crohn's disease.[10] It is therefore crucial to (1) identify if a patient has only 1 disease or both and (2) detect whether the condition affects other members of the family. Finally, some diseases can affect multiple organs, and clinical presentation often includes a wide array of symptoms; for example, lupus enteritis is suspected when a patient presents with abdominal pain, diarrhea, and vomiting.[11] Given the low incidence (only 5% of lupus patients have diarrhea,[12] and the incidence of lupus is about 5.5/100 000 persons[13]) and nonspecific clinical findings, it is important to identify those cases.

We developed a method that detects the negated segments and the segments related to the family history in the narrative reports. This method is part of the Extract-Transform-Load process of the CDW (Figure 1). We evaluated the benefits of using this method to distinguish between true positives (TPs) and false positives (FPs) in 3 corpora of rare disease patient records related, respectively, to Crohn's disease, lupus, and NPHP1 from the CDW.

RELATED WORK

Negation detection

Most of the published NLP methods are based on concept extraction and their assertion classifications.[14–17] The most common algorithm for negation detection is NegEx.[18] It is based on regular expressions targeted before and after a concept in a document to determine whether a concept is negated. NegEx has been ported and evaluated on clinical texts in French and has shown good performance (a recall of 85% and a precision of 89%) on cardiology notes.[19,20]

In 2006, Goryachev et al.[21] evaluated NegEx and 3 other methods of negation detection. Among them, NegExpander was developed by Aronow et al.[22] and was based on NegEx, with extension to conjunctive phrases and to all Unified Medical Language System (UMLS) terms. The 2 other methods were based on machine-learning classifiers using Weka machine-learning software and a naïve Bayes classifier associated with a support vector machine classifier. Goryachev concluded that NegEx was the most effective.

Family history detection

As for family history, most authors have developed methods that extract family information from specific or focused sections in clinical reports. Friedlin and McDonald[23] focused on sentences identified by titles with a family history–type phrase in the admission notes. They used the Regenstrief data eXtraction tool to associate 12 diseases to family histories in these sentences, with a precision and recall of 97% and 96%, respectively. Goryachev et al.[24] also assumed that the medical narrative reports were structured in sections. Inside each section identified using section headings, the entities were extracted and categorized using the UMLS semantic types. The authors developed a set of rules to associate the finding with a family member or with the patient. Their pipeline achieved 97.2% sensitivity and 99.7% specificity in detecting the family history findings. Lewis et al.[25] examined the entire text, considering that a patient's family history may be spread throughout the patient's clinical records and is mostly recorded in the clinical notes. They used a Stanford NLP

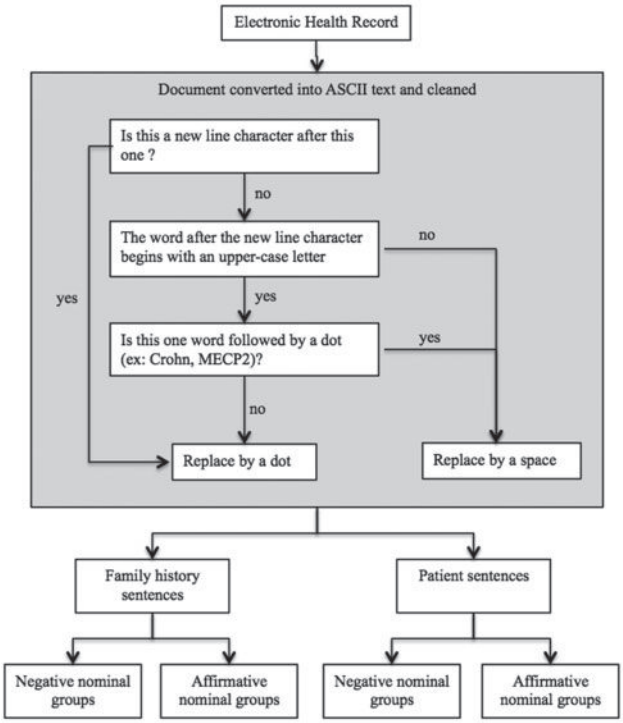


Figure 2. Overview of the pipeline.

Parser to detect dependency between the disease and the family relationship with a precision of 61% and a recall of 51%.

The ConText algorithm developed by Chapman et al.[26] is based on the NegEx approach; it is a regular-expression–based algorithm that searches for trigger terms preceding or following the indexed clinical conditions to determine if clinical conditions mentioned in the clinical reports are negated, hypothetical, historical, or experienced by someone other than the patient. Chapman et al. showed that it was relevant to use only regular expressions to detect these contextual values to classify the concepts extracted from the text.

The main limitations of these approaches are the restriction to preidentified concepts and/or the use of semistructured reports (reports with explicit sections).

Our aim is to extract subtexts from each original patient record and classify them into 4 categories: patient–not negated/patient–negated/family history–not negated/family history–negated. These subtexts would be integrated through the Extract-Transform-Load process in the CDW to improve the performance of the full-text search engine.

METHODS

Overview

We built a pipeline (Figure 2) to process the medical records in 3 steps: cleaning text, detecting context (patient data/family history), and detecting negation in each context. This pipeline was designed to process unstructured reports and free text from multiple sources and authors.

Text processing

The first step consisted of converting all the Word® or PDF (Portable Document Format) documents to ASCII (American Standard Code for Information Interchange), then cleaning the documents from erroneous newline characters generated by the automated conversion. To determine whether to replace a newline character, we developed a decision tree based on regular expression rules to choose the correct character replacement for each newline character (Figure 2).

Family history detection

We listed all the French words related to family relationships (Supplementary Appendix 1). The algorithm takes into account the age of the patient at the date of the document. For example, when the patient is under 18 years old, it does not consider the words "son" and "daughter" in the list because this is the parent referring to their child, who is actually the patient. Moreover, we added several expressions that were meaningful for family history such as "family history," "paternal ancestry," and "maternal ancestry."

Each sentence in which a term of the above list matched was classified as belonging to the family history context; otherwise, the sentence was classified in the patient context.

Negation detection

We built a corpus of 3900 narrative documents by querying Dr Warehouse® for 4 diseases: terminal renal failure, autism, Rett Syndrome, and Currarino.

We used this corpus to build (Table 1):

- a list of French regular expression rules to split the sentence into propositions or nominal groups (Supplementary Appendix 2).
- a list of negated regular expressions to determine whether a proposition had a negative meaning.
- a list of exclusion rules for the double negatives such as "we cannot exclude" or "without any doubt."

The algorithm classifies the expression of normality as negative information, e.g., "gene MECP2 normal" is equivalent to "MECP2 not mutated." We decided to keep hypothetical diagnostic and "research for" as affirmative information as long as no final diagnosis had been confirmed. Based on these rules, the algorithm classifies each proposition and nominal group as negative or non-negative.

CDW integration

Dr Warehouse® relies on Oracle® 11g, and the search engine is based on the Oracle® Text module. For a given document, all of the triplets {subtext, context, negation} are stored in Dr Warehouse®. Context is either "patient" or "family history."

EVALUATION

To evaluate the impact of negation and family history detection on the performance of the search engine, we used the pipeline on the entire CDW. The textual documents were de-identified using an internally developed algorithm based on name, first name, birth date, address, and hospital ID. To protect confidentiality, authorized internal staff at the Necker Hospital conducted the study.

We evaluated the system on 3 corpora extracted from Dr Warehouse® at Necker/Imagine Institute by querying the whole CDW, which contains 1.6 million EHRs for 350 000 individual patients (January 2016). The corpora corresponded to 3 clinical use cases, namely "Crohn's and diabetes," "lupus and diarrhea," and "NPHP1." For each of them, we counted the number of TPs, FPs, true negatives (TN), and false negatives (FNs) without any filtering

Table 1. List of negated regular expressions and exclusion rules

	Negated expressions	Exclusion rules
1	/[^a-z]pas\s[a-z]*\s*d/i	/[^a-z]pas\s*([a-z]*\s){0,2}doute/i /[^a-z]pas\s*([a-z']*\s*){0,2}eliminer/i /[^a-z]pas\s*([a-z']*\s*){0,2}exclure/i /[^a-z]pas\s*([a-z']*\s*){0,2}probleme/i /[^a-z]pas\s*([a-z']*\s*){0,2}soucis/i /[^a-z]pas\s*([a-z']*\s*){0,2}objection/i /\sne reviens\s+pas/i
2	/\sn(e|')(l[ae]|l')?[a-z]+pas[^a-z]/i	/[^a-z]pas\s*([a-z]*\s){0,2}doute/i /[^a-z]pas\s*([a-z']*\s*){0,2}eliminer/i /[^a-z]pas\s*([a-z']*\s*){0,2}exclure/i /[^a-z]pas\s*([a-z']*\s*){0,2}probleme/i /[^a-z]pas\s*([a-z']*\s*){0,2}soucis/i /[^a-z]pas\s*([a-z']*\s*){0,2}objection/i /\sne reviens\s+pas/i
3	/[^a-z]sans\s/i	/[^a-z]sans\s+doute/i /[^a-z]sans\s+probleme/i /[^a-z]sans\s+soucis/i /[^a-z]sans\s+objection/ /[^a-z]sans\s+difficult/
4	/aucun/	/aucun\s+doute/i /aucun\s+probleme/i /aucun\s+soucis/i /aucune\s+objection/
5	/\selimine/i	/[^a-z]pas\s+d'eliminer/i /[^a-z]sans\s+eliminer/i
6	/\sinfirme/i	/[^a-z]pas\s+d'infirmer/i /[^a-z]sans\s+infirmer/i
7	/[^a-z]exclu[e]?[s]?[^a-z]/i	/pas\s+d'exclure/i /\spas\s+exclu[\s]/i
8	/[^a-z]jamais\s[a-z]*\s*d/i	
9	/[^a-z]ni\s/i	
10	/oriente pas vers/i	
11	/:\s*non[^a-z]/i	
12	/^\s*non[^a-z]+$/i	
13	/:\s*aucun/i	
14	/:\s*exclu/i	
15	/:\s*absent/i	
16	/:\s*inconnu/i	
17	/absence/i	
18	/absent/i	
19	/\sne pas\s/i	
20	/\snegati.*/i	
21	/[^a-z]normale?s?\s/i	/pas\s+[^a-z]normale?s?\s/i
22	/[^a-z]normaux\s/i	/pas\s+[^a-z]normaux\s/i

Suggested English equivalent: do not (1, 2, 19), without (3), no/none (4, 11, 12, 13), eliminate (5, 6), exclude/exclusion (7, 14), never (8), neither (9), not lead to (10), absent/absence (15, 17, 18), unknown (16), negative/negation (20), normal (21, 22).

(i.e., before applying our algorithm), with family history detection only, with negation detection only, and with both detection algorithms. We evaluated the FNs regarding negation and family history detection only, as it was not possible to calculate them for the non-filtered queries. We defined TPs as patients with all terms present in their EHRs in a non-negative expression and with no family history expression. The TN patients were defined as having all the terms but at least 1 in a negative expression or in a family history context.

Two persons independently evaluated the system and the Kappa score was calculated. In case of discordance, a consensus was reached. Four metrics were used to assess the performance of our algorithm: recall, precision, specificity, and F-measure.

RESULTS

The overall interevaluator agreement measured by the Kappa coefficient was 0.98. Table 2 displays the results for each use case. Before applying our algorithm, 145 patients (262 documents) were classified as "Lupus and diarrhea," 173 patients (269 documents) "Crohn's and diabetes," and 32 patients (95 documents) "NPHP1." This corresponds to a total of 626 heterogeneous documents distributed on several clinical services (Supplementary Appendix 3) and several types of records (Supplementary Appendix 4). For the 3 use cases, negation detection and family history filters provided, separately, an increased precision and F-measure. The combination of

Table 2. Precision, recall, specificity and F-measure for each use case and for pooled data

Lupus and diarrhea	TP	FP	TN	FN	Precision	Recall	Specificity	F-measure
No filtering	53	92	0	0	0.37	1.00	0.00	0.54
Family history	53	41	51	0	0.56	1.00	0.55	0.72
Negation	52	39	53	1	0.57	0.98	0.58	0.72
Family history and negation	52	5	87	1	0.91	0.98	0.95	0.95
Crohn's and diabetes								
No filtering	23	150	0	0	0.13	1.00	0.00	0.23
Family history	22	25	125	1	0.47	0.96	0.83	0.63
Negation	23	107	43	0	0.18	1.00	0.29	0.30
Family history and negation	22	8	142	1	0.73	0.96	0.95	0.83
NPHP1								
No filtering	21	11	0	0	0.66	1.00	0.00	0.79
Family history	21	11	0	0	0.66	1.00	0.00	0.79
Negation	21	4	7	0	0.84	1.00	0.64	0.91
Family history and negation	21	4	7	0	0.84	1.00	0.64	0.91
Total								
No filter	97	253	0	0	0.28	1.00	0.00	0.43
Family history	96	77	176	1	0.55	0.99	0.70	0.71
Negation	96	150	103	1	0.39	0.99	0.41	0.56
Family history and negation	95	17	236	2	0.85	0.98	0.93	0.91

Table 3. For each report type, number of propositions (prop) with negative regex over the total number of propositions, and the number of sentences with family regex over the total number of sentences

Report type	Nb of prop with negative regex over the total nb of prop (%)	Nb of sentences with family regex over the total nb of sentences (%)
Consultation reports	2659/17296 (15)	683/9346 (7)
Day hospitalization reports	1413/11253 (13)	280/7448 (4)
Emergency	32/144 (22)	2/90 (2)
Hospitalization reports	9262/69861 (13)	1074/45773 (2)
Letter	1618/13200 (12)	196/8077 (2)
Operative reports	26/267 (10)	4/182 (2)

both filters further improved the precision score and F-measure. The precision increased from 0.37 to 0.91 for "Lupus and diarrhea," from 0.13 to 0.73 for "Crohn's and diabetes," and from 0.66 to 0.84 for "NPHP1." For "lupus and diarrhea" and "Crohn's and diabetes," the recall decreased to, respectively, 0.98 and 0.96 with the filters. For "NPHP1," the recall remained equal to 1.

The overall precision, recall, specificity, and F-measure for the pooled data were 0.85, 0.98, 0.93, and 0.91, respectively (Table 2). In addition, only 2 patients were erroneously excluded out of the 95 TPs by the algorithm.

While negative propositions were more often present in emergency reports, with 22% of the propositions classified as negative, family history sentences were more present in consultation reports (Table 3).

The regex negative expressions most useful in excluding FP patients were rules 1, 2, and 9, with, respectively, 64, 16, and 24 FP patients detected. The only FN patient induced was due to the rule "absence of" (Supplementary Appendix 5).

The processing (combining negation and family history) took an average of 18 ms per document. The queries used to build the corpora took, respectively, 3 s for "lupus and diarrhea," 1 s for "Crohn's and diabetes," and 1 s for "NPHP1" on a server with 32 Go RAM and 8 cores 2.4 Ghz.

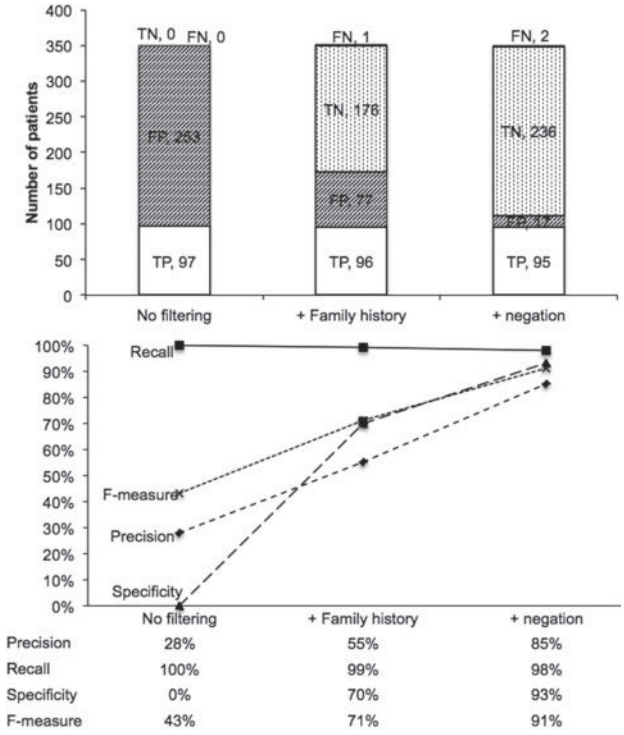
DISCUSSION

The aim of the present study was to develop an automated tool to filter negated expressions and family context from clinical narrative reports. The results indicate that use of the automated approach is feasible and dramatically decreases the rate of FPs: 71% before filtering vs 15% using our algorithm. Our method achieved very good overall precision (0.85), recall (0.98), specificity (0.93), and F-measure (0.91) and also for each use case (Figure 3). One of the strengths of this study is that these results were obtained on a comprehensive corpus that contains both inpatient and outpatient reports.

Limitations and perspectives

Three major causes explain the remaining FPs (17 patients):

- Seven were due to an incorrect split of a sentence in nominal groups. This error is due to undeleted erroneous newline characters.
- Five were misclassified because a diagnosis was disscussed but was ultimately deferred until a later time.
- Our system does not resolve issues of coreference (5 cases), which is the task of finding all expressions that refer to the same



	No filtering	+ Family history	+ negation
Precision	28%	55%	85%
Recall	100%	99%	98%
Specificity	0%	70%	93%
F-measure	43%	71%	91%

Figure 3. Both the number of patients retrieved and the accuracy (represented by precision, recall, specificity, and F-measure) are presented without filtering, with family history detection and with both detection (family history and negation).

entity in a text, i.e., "The father is Caucasian. He has Crohn's disease." This type of limitation is also found in the other systems mentioned in the Related work section.[24] There is ongoing research on computational methods for coreference resolution that may be implemented in our system in the future.[27,28]

Regarding the evaluation of certainty, we decided to use only 2 modalities: non-negative and negative. A third level should be added to our scale to consider suspicion.

Comparison to other works

Harkema et al.[26] applied ConText to 6 types of clinical reports, e.g., Surgical Pathology, in which the conditions experienced by someone other than the patient were very rarely found. The number of occurrences of "other experiencer" occurred only 5 times in all the reports in their evaluation set; i.e., it was strictly 0% for radiology reports, surgical pathology, and operative procedures. Conversely, 24% of our report set corresponding to rare disease patients contained mention of some condition experienced by their family members.

Tanushi highlighted the lack of regular expressions representing double negation or normality in Negex.[29] Similarly, we did not find any of them in the csv file displayed by Chapman.[19] Noticeably, we identified 497 occurrences of double negatives with "we cannot exclude" and 7831 occurrences of "without any doubt" in the CDW. We proposed a list of expressions to take into account potential misclassification due to double negation and normality.

Deleger developed an algorithm to detect negation of medical problems in cardiology notes in French.[20] Their algorithm focused on negation detection and their method was based on concept classification (F-measure 0.87). Our algorithm combines detection of family history and negation, and was applied to a large CDW of 1.6 million EHRs. The algorithm is part of a full-text search engine with the objective of classifying patients as cases or not (F-measure 0.91).

CONCLUSION

We developed an integrated pipeline to enable negation and family history context detection in the full-text search engine of a document-oriented CDW. The automated method achieved an overall F-measure of 0.91. The method is generalizable, and can be adapted to English and other languages.

The tool is available at https://github.com/imagine-bdd/DrWH-negation

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

Substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work: *Conception of the work*: NG, AB. *Acquisition of data*: NG, VB. *Interpretation of data:* NG, AN, RS, AB. *Evaluation:* RS, AB. Drafting the work or revising it critically for important intellectual content: *Drafting the work*: NG, AN, AB. *Revising the work critically for important intellectual content*: NG, AN, VB, RS, AB. Final approval of the version to be published: NG, AN, VB, RS, AB. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: NG, AN, VB, RS, AB.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

REFERENCES

1. Murphy SN, Mendis ME, Berkowitz DA, *et al*. Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu Symp Proc*. 2006;2006:1040.
2. Hebbring SJ, Rastegar-Mojarad M, Ye Z, *et al*. Application of clinical text data for phenome-wide association studies (PheWASs). *Bioinformatics*. 2015;31(12):1981–7.
3. Cuggia M, Garcelon N, Campillo-Gimenez B, *et al*. Roogle: an information retrieval engine for clinical data warehouse. *Stud Health Technol Inform*. 2011;169:584–8.
4. Huan Mo, William K Thompson, Luke V Rasmussen, *et al*. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J Am Med Inform Assoc*. 2015;22(6):1220–30.
5. Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput Biol* 2012;8(12):e1002823.
6. Cuggia M, Bayat S, Garcelon N, *et al*. A full-text information retrieval system for an epidemiological registry. *Stud Health Technol Inform*. 2010;160(Pt 1):491–5.

7. Escudié JB, Rance B. Reviewing 741 patients records in two hours with FASTVISU. *Proc AMIA Symp*. 2015; 2015:553–9.
8. Chapman WW, Bridewell W, Hanbury P, *et al*. Evaluation of negation phrases in narrative clinical reports. *Proc AMIA Symp*. 2001;2001:105–9.
9. Chu D, Dowling JN, Chapman WW. Evaluating the effectiveness of four contextual features in classifying annotated clinical conditions in emergency department reports. AMIA Annu Symp Proc. 2006;2006:141–5.
10. Sharp RC, Abdulrahim M, Naser ES, Naser SA. Genetic Variations of PTPN2 and PTPN22: role in the pathogenesis of type 1 diabetes and Crohn's disease. *Front Cell Infect Microbiol*. 2015;5:95.
11. Sran S, Sran M, Patel N, Anand P. Lupus enteritis as an initial presentation of systemic lupus erythematosus. *Case Rep Gastrointest Med*. 2014;2014:962735.
12. Sultan SM, Ioannou Y, Isenberg DA. A review of gastrointestinal manifestations of systemic lupus erythematosus. Rheumatology *(Oxford)*. 1999; 38(10):917–32.
13. Somers EC, Marder W, Cagnoli P, *et al*. Population-based incidence and prevalence of systemic lupus erythematosus: the Michigan Lupus Epidemiology and Surveillance program. *Arthritis Rheumatol*. 2014; 66(2):369–78.
14. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507–13.
15. South BR, Phansalkar S, Swaminathan AD, *et al*. Adaptation of the NegEx algorithm to Veterans Affairs electronic text notes for detection of influenza-like illness (ILI). *AMIA Annu Symp Proc*. 2007;2007:1118.
16. Skeppstedt M. Negation detection in Swedish clinical text: an adaption of NegEx to Swedish. *J Biomed Semantics*. 2011;2 (Suppl 3):S3.
17. King B, Wang L, Provalov I, *et al*. Cengage learning at TREC 2011 Medical Track. In *National Institute of Standards and Technology*. *Proc TREC*. 2011. Maryland.
18. Chapman W, Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001;34(5):301–10.
19. Chapman WW, Hillert D, Velupillai S, *et al*. Extending the NegEx lexicon for multiple languages. *Stud Health Technol Inform*. 2013;192:677–81.
20. Deléger L, Grouin C. Detecting negation of medical problems in French clinical notes. *In* Proc the 2nd ACM SIGHIT International Health Informatics Symposium *(IHI '12)*. New York, NY, USA: ACM; 2012: 697–702.
21. Goryachev S, Sordo M, Zeng QT, *et al*. Implementation and evaluation of four different methods of negation detection. *Boston, MA, DSG*. 2006. Boston, Maryland; Decision Systems Group Technical Report.
22. Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. *J Am Med Inform Assoc*. 1999; 6(5):393–411.
23. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. *AMIA Annu Symp Proc*. 2006;2006:925.
24. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. *AMIA Annu Symp Proc*. 2008; 2008:247–51.
25. Lewis N, Gruhl D, Yang H. Extracting family history diagnosis from clinical texts. *Int Conf Bioinform Comput Biol*. New Orleans, Louisiana. 2011: 128–133.
26. Chapman W, Chu D, Dowling JN. ConText: an algorithm for identifying contextual features from clinical text. *In Proc Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing (BioNLP '07)*. USA, PA: Association for Computational Linguistics, Stroudsburg; 2007: 81–88.
27. Kim Y, Riloff E, Gilbert N. The taming of reconcile as a biomedical conference resolver. ACL. *Workshop BioNLP- Shared task;Proceedings of BioNLP Shared Task 2011 Workshop*, Portland, Oregon, USA, Association for Computational Linguistics June 2011;2011:89–93.
28. Zheng J, Chapman W, Miller T, *et al*. A system for coreference resolution for the clinical narrative. *Am Med Inform Assoc*. 2012;19(4):660–7.
29. Tanushi H, Dalianis H, Duneld M, *et al*. Negation scope delimitation in clinical text using three approaches: NegEx, PyConTextNLP and SynNeg. In: S Oepen, K Hagen, J Bondi Johannessen, eds. *Proc 19th NODALIDA, NEALT*. Linköping, Sweden: Linköping University Electronic Press; 2013:387–97.

3.2.3. Enrichissement terminologique

Nous avons utilisé l'UMLS Metathesaurus comme dictionnaire afin de réaliser l'enrichissement terminologique. La version 2017AA de l'UMLS Metathesaurus contient l'alignement de 201 terminologies médicales en 25 langues pour un total de 13,5 millions de termes (3,5 millions de concepts) ("Statistics - 2017AA Release," n.d.).

Nous avons développé deux méthodes distinctes pour les concepts phénotypiques et les concepts « génétiques » (correspondant à des gènes). Les concepts phénotypiques sont définis par les types sémantiques suivants : « Sign or Symptom », « Disease or Syndrome », « Finding », « Pathologic Function », « Congenital Abnormality », « Physiologic Function », « Anatomical Abnormality », « Neoplastic Process », « Acquired Abnormality », « Mental or Behavioral Dysfunction » et qui ne sont pas dans le type sémantique « Gene or Genome ».

Nous n'avons pas utilisé le groupe sémantique « disorder » (McCray et al., 2001), car il contient aussi les types sémantiques « Injury or Poisoning », « Experimental Model of Disease » et « Cell or Molecular Dysfunction » qui ne sont pas des phénotypes à proprement parler.

Les concepts génétiques sont définis par le type sémantique « Gene or Genome ».

Extraction phénotypique

Nous avons défini un sous corpus de l'UMLS contenant tous les termes français dans les types sémantiques phénotypiques. Nous avons éliminé les concepts de moins de trois caractères pour réduire le nombre d'erreurs possibles de faux positifs. A partir des 13 518 566 termes (3 465 486 concepts) de l'UMLS ("Statistics - 2017AA Release," n.d.), nous en conservons 93 366 (51 930 concepts).

Chaque terme du sous corpus est transformé en expression régulière dans laquelle les caractères non alphanumériques sont remplacés par [^a-z0-9]*. Les caractères accentués sont remplacés par les caractères équivalents non accentués. Les « e » en fin de mot sont remplacés par « es? » ce qui permet de prendre en compte les dérivées pluriels des termes dans le texte.

Les termes du sous corpus sont testés par ordre décroissant de longueur afin de conserver la granularité la plus fine à chaque fois. Pour chaque triplet *texte-contexte-certitude* contenu dans la table DWH_TEXT, on remplace les caractères accentués par les caractères équivalents non accentués. Puis on teste la présence de chaque expression régulière des termes du sous corpus dans le texte. Le test réalisé est insensible à la casse. Si l'expression régulière renvoie une correspondance, le terme est supprimé du texte afin d'éviter la détection d'un terme inclus dans celui ci (par exemple, si le terme « insuffisance rénale terminale » est retrouvé, le terme « insuffisance rénale » ne sera pas retrouvé). On conserve le terme retrouvé, le concept correspondant et le nombre d'occurrences trouvées dans le texte. Ces données sont enregistrées

dans la table DWH_ENRSEM avec le contexte, le niveau de certitude (négation, non négation) et la clé du document correspondant au texte analysé.
Le fait de conserver le texte permet de calculer le terme préféré par concept par rapport aux usages des médecins de l'hôpital. Le terme préféré est indiqué dans la table DWH_THESAURUS_ENRSEM. Nous pourrons ainsi afficher les termes utilisés par les médecins.

Extraction des gènes

Nous avons créé un sous corpus de l'UMLS contenant les concepts appartenant au type sémantique « Gene or Genome » et présents dans les thesaurus HUGO ou OMIM. Nous excluons les termes de moins de 3 caractères et de plus de 55 caractères (Wu et al., 2012). Nous en conservons 152 722 (40 751 concepts).
Nous effectuons les mêmes remplacements des caractères non alphanumériques par [^a-z0-9]*. Si le terme contient un caractère numérique, l'expression régulière est insensible à la casse. Si le terme ne contient pas de caractère numérique, l'expression régulière est transformée en majuscule et est sensible à la casse (Groza and Verspoor, 2015). Cela permet de réduire les ambiguïtés possibles. L'algorithme reprend ensuite le même processus que pour l'extraction phénotypique. Les données sont aussi enregistrées dans la table DWH_ENRSEM avec le contexte, le niveau de certitude (négation, non négation) et l'id du document correspondant au texte analysé.

Thesaurus des concepts extraits

Le corpus DWH_THESAURUS_ENRSEM est construit à partir des concepts UMLS reconnus. On calcule le nombre de patients par concept, le nombre de documents par terme, et le nombre de patients pour le concept et ses concepts fils. Ces fréquences sont ensuite utilisées pour les calculs de TF-IDF. La hiérarchie est créée à partir de la table MRHIER de l'UMLS (McInnes et al., 2009), cette hiérarchie correspond aux relations « is a » entre deux concepts. Si un concept appartient à plusieurs hiérarchies, nous les conservons toutes sans a priori. Nous ajoutons si nécessaire les concepts parents s'ils sont absents du sous corpus d'origine. Si le concept parent n'a pas de terme en français, nous prenons le terme anglais. La table DWH_THESAURUS_ENRSEM_GRAPH est créée à partir de la table DWH_THESAURUS_ENRSEM.

Nous avons intégré dans l'entrepôt de données les données de Gene Ontology disponibles sur le serveur FTP d'Entrez Gene ("Home - Gene - NCBI," n.d.) :

- Homo_sapiens.gene_info qui est un sous ensemble des gènes disponibles réduit aux gènes humains. Il permet de lister tous les synonymes des gènes.

- gene2go qui nous permet d'associer aux gènes les processus biologiques, les composants cellulaires et les fonctions moléculaires décrits dans Gene Ontology.

Ces données ne sont pas intégrées dans l'enrichissement des documents. Nous utiliserons ce thesaurus uniquement pour agréger les patients par les processus biologiques, les composants cellulaires et les fonctions moléculaires mis en jeu par les gènes mutés chez ces patients.

L'enrichissement terminologique

L'expansion d'une requête est décrite comme l'explosion sémantique des termes utilisés (Plovnick and Zeng, 2004). L'UMLS Metathesaurus est utilisé pour enrichir la requête avec les synonymes et les hyponymes retrouvés par la relation de type « is a » (Ding et al., 2007; Grabar et al., 2008; Griffon et al., 2012; Martinez et al., 2014). Notre approche diffère des approches précédentes, parce qu'au lieu d'enrichir la requête nous enrichissons l'indexation des textes. L'enrichissement du document pour améliorer la recherche d'information a été décrit et évalué par (Müller et al., 2010) et par (Köhncke et al., 2013). Müller montre une meilleure Mean Average Precision (MAP) en utilisant l'enrichissement du texte (MAP=0.231) en comparaison avec l'expansion de la requête (MAP=0.218). Köhncke améliore en moyenne de 74.5% la précision du moteur de recherche en enrichissant le texte.

Pour réaliser cet enrichissement, nous concaténons dans la colonne ENRICH_TEXT de la table DWH_TEXT le texte contenu dans la colonne TEXT, les libellés des concepts extraits depuis ce texte (libellé préféré et les synonymes), et tous les libellés de leurs concepts parents (en utilisant la hiérarchie enregistré dans DWH_THESAURUS_ENRSEM_GRAPH correspondant à la relation « is a » de l'UMLS).

Ainsi un texte qui contient le terme « diabète » aura le concept UMLS extrait C0011849 (libellés : *diabète et diabète sucré*), qui a pour parent C0851431 (libellé : *Troubles du métabolisme du glucose*), qui a lui même pour parent C0014130 (libellés : *Affections endocriniennes, Trouble endocrinaire, Trouble endocrinien, Endocrinopathies, Maladies des glandes endocrines, Maladies du système endocrinien, Maladies endocriniennes*). On aura donc dans le texte enrichi :

Le texte d'origine ; diabète ; diabète sucré : Troubles du métabolisme du glucose ; Affections endocriniennes ; Trouble endocrinaire ; Trouble endocrinien ; Endocrinopathies ; Maladies des glandes endocrines ; Maladies du système endocrinien ; Maladies endocriniennes

La requête en texte libre « trouble endocrinien » renverra donc les documents qui contiennent « diabète ».

Pour les parties de texte qui sont considérées comme des négations, nous n'ajoutons pas les concepts parents, mais les concepts fils. Car l'absence d'un symptôme ne signifie pas l'absence des symptômes parents, mais bien l'absence des symptômes fils.

3.3. Phénotypage haut débit

eMERGE (Electronic Medical Records for Genetic Research) Consortium a publié en 2011 l'intérêt d'utiliser les données cliniques issues du dossier patient pour identifier de nouvelles associations phénotypiques (Kho et al., 2011). Lors de l'enrichissement terminologique nous avons extrait l'ensemble des concepts phénotypiques des comptes rendus cliniques. L'objectif est d'identifier les phénotypes les plus pertinents dans un groupe de patients identifiés. Dans l'entrepôt de données de Necker, la requête « syndrome de Rett » renvoie environ 300 patients. L'extraction phénotypique de leurs comptes rendus donne environ 1300 concepts distincts (en dehors des concepts classés comme négation ou antécédents familiaux). Afin de déterminer les concepts les plus pertinents, nous avons calculés la fréquence des patients pour chaque concept dans la population étudiée, le score TF-IDF des concepts (Jones, 1972) permettant de prendre en compte la fréquence du concept dans l'ensemble de l'entrepôt de données et la fréquence de ce concept dans la population étudiée. Le résultat est un tri des concepts suivant ces différentes métriques permettant de visualiser les Top50 concepts.

Nous avons évalué notre capacité à retrouver des phénotypes pertinents sur 6 cohortes de patients : Syndrome de Rett, Syndrome PI3K Delta activé, Syndrome de Lowe, Syndrome de Silver Russell, Déficit en Dock8, Syndrome de Bardet Biedl. Nous avons comparé nos résultats avec Orphadata (INSERM, 1997) qui propose une description phénotypique des maladies rares avec le thesaurus HPO. Notons que le syndrome PI3K Delta activé n'est pas encore décrit dans Orphanet à l'heure où nous rédigeons ce manuscrit. Nous avons par ailleurs réalisé une évaluation par les médecins experts de l'hôpital Necker-Enfants Malades de chaque cohorte. Ils ont évalué les 50 premiers concepts retrouvés par les deux scores séparément.

Si le recouvrement avec Orphadata reste honorable il reste des difficultés liées à la structure de l'UMLS et à son manque de vocabulaire français. Nous avons une excellente évaluation par les experts. Nous retrouvons par ailleurs des concepts absents d'Orphadata, validés par les experts et bien présents dans la littérature : par exemple, l'ostéoporose pour le syndrome de Rett.

Nous avons soumis cet article à Orphanet Journal of Rare Diseases en septembre 2017.

Next Generation Phenotyping using narrative reports in a rare disease Clinical Data Warehouse

Nicolas Garcelon[1,2], Antoine Neuraz[2,3], Rémi Salomon[1,4], Nadia Bahi-Buisson[1,5], Jeanne Amiel[1,6,7], Capucine Picard[1,8,9], Nizar Mahlaoui[1,8,10,11], Vincent Benoit[1], Anita Burgun[2,3,12], Bastien Rance[2,12]

[1]Institut Imagine, Paris Descartes Paris Descartes-Sorbonne Paris Cité University, Paris, France;
[2]Institut National de la Santé et de la Recherche Médicale (INSERM), Centre de Recherche des Cordeliers, UMR 1138 Equipe 22, Paris Descartes, Sorbonne Paris Cité University, Paris, France;
[3]Department of Medical Informatics, Necker-Enfants Malades Hospital, Assistance Publique des Hôpitaux de Paris (AP-HP), Paris, France
[4]Pediatric Nephrology, Necker Enfants Malades Hospital AP-HP, Université Paris Descartes, Paris, France;
[5]Pediatric Neurology, Necker Enfants Malades Hospital AP-HP, Université Paris Descartes, Paris, France;
[6]Laboratory of embryology and genetics of congenital malformations, INSERM UMR 1163, Institut Imagine, Paris, France;
[7]Department of Genetic, Necker Enfants Malades Hospital AP-HP, Université Paris Descartes, Paris, France
[8] Laboratory of Lymphocyte Activation and Susceptibility to EBV infection, INSERM UMR 1163, Paris, France, Paris Descartes Sorbonne Paris Cité University, Imagine Institute, Paris, France
[9]Study center for primary immunodeficiencies (CEDI) Necker Enfants Malades Hospital AP-HP, Université Paris Descartes, Paris, France
[10] French National Reference Center for Primary ImmunoDeficiencies (CEREDIH), Necker Enfants Malades Hospital AP-HP, Université Paris Descartes, Paris, France ;
[11]Pediatric Immuno-Haematology and Rheumatology Necker Enfants Malades Hospital AP-HP, Université Paris Descartes, Paris, France ;.
[12]Hôpital Européen Georges Pompidou, AP-HP, Université Paris Descartes, Sorbonne Paris Cité, France

Corresponding author:
Nicolas Garcelon
Imagine - Institute for genetic diseases
24 boulevard du Montparnasse
75015 Paris
France
Email: nicolas.garcelon@institutimagine.org
Tel: 0033.1.42.75.44.57

Abstract

Introduction: Secondary use of data collected in Electronic Health Records opens perspectives for increasing our knowledge of rare diseases. The clinical data warehouse (named Dr Warehouse) at the Necker-Enfants Malades Children's Hospital contains data collected during normal care for thousands of patients. Dr Warehouse is oriented toward the exploration of clinical narratives. In this study, we present our method to find phenotypes associated with diseases of interest.
Method: We leveraged the frequency and TF-IDF to explore the association between clinical phenotypes and rare diseases. We applied our method in six use cases: phenotypes associated with the Rett, Lowe, Silver Russell, Bardet-Biedl syndromes, DOCK8 deficiency and Activated PI3-kinase Delta Syndrome (APDS). We asked domain experts to evaluate the relevance of the top-50 (for frequency and TF-IDF) phenotypes identified by Dr Warehouse and computed the average precision and mean average precision.
Results: Experts concluded that between 16 and 39 phenotypes could be considered as relevant in the top-50 phenotypes ranked by descending frequency discovered by Dr Warehouse (resp. between 11 and 41 for TF-IDF). Average precision ranges from 0.55 to 0.91 for frequency and 0.52 to 0.95 for TF-IDF. Mean average precision was 0.79. Our study suggests that phenotypes identified in clinical narratives stored in Electronic Health Record can provide rare disease specialists with candidate phenotypes that can be used in addition to the literature.
Keywords: Data warehouse, next generation phenotyping, data mining, rare diseases, natural language processing

1. Introduction

The global trend toward digital health in the US and in Europe has led to an unprecedented adoption of Electronic Health Records (EHRs). By the end of 2014, 83% of US physicians [1] and 75% of hospitals [2] used some form of EHRs. The increasing number of EHRs opens strong perspectives for the secondary use of data collected during the care process. Many hospitals are now equipped with Clinical Data Warehouses (CDW) integrating all the data produced during the care of the patients for research purposes [3–5]. CDWs gather a large variety of information, ranging from structured data (e.g. diagnosis codes, laboratory test results,…) to free-text clinical narratives and images. Structured data include coded data using terminologies like the International Classification of Diseases, and questionnaires that provide precise, standardized but somehow limited information. Conversely free-text reports are produced without constraints and may be used to express nuanced, unexpected, and unexplained signs or symptoms regarding the patient case. Clinical narratives collect information from all aspects of the patient care that might not be collected anywhere else in clinical information system including history of the disease, family history, fine-grained description of all the symptoms, hypothesis of diagnosis or treatment, information from treatments received outside of the hospital, and so forth. Previous studies in different contexts showed the importance of free-text in EHRs. For example Raghavan *et al.* identified that unstructured data were essential to solve trial criteria from two studies. [6]. The value of text data is even more important to detect phenotypes in specialized hospitals treating patients with rare diseases and for outpatients, for whom clinical information is barely coded [7].

Rare diseases represent a large group of heterogeneous conditions and some cases remain undiagnosed for a long time. A precise phenotypic description of such diseases can be problematic given the small number of cases and the heterogeneity of the phenotypes. Leveraging large CDWs, could be helpful to enrich this description. While structured (standardized) questionnaires exist for several rare diseases (e.g., in France [8,9]), part of the clinical description is still present only in free text in EHRs. We hypothesized that mining large collections of clinical texts in hospitals specialized in rare diseases could offer interesting perspectives to enrich the descriptions provided by dedicated knowledge bases. We investigated this hypothesis at the *Necker Enfants Malades Hospital* (Necker Children Hospital), a children's hospital in Paris that is associated with the *Imagine* research institute, specialized in genetic diseases, and hosts 15 national reference centers for rare diseases. We illustrate our approach on six rare diseases: DOCK8 deficiency, the Activated PI3-kinase Delta Syndrome (APDS), Rett, Lowe, Silver Russell and Bardet Biedl syndromes. The combined immunodeficiency due to DOCK8 deficiency (prevalence less than 1/1,000,000) is a form of autosomal recessive combined immunodeficiency (T, B and NK cells), characterized by recurrent lung infections, cutaneous viral infections, allergy, severe skin inflammation and susceptibility to cancer with a high level of IgE [10]. DOCK8 deficiency is caused by homozygous or compound heterozygous mutations in *DOCK8* gene [11].

The activated phosphoinositide 3-kinase-δ (PI3Kδ) syndrome (APDS) (estimated prevalence < 1 /1,000,000) is characterized by immunodeficiency and recurrent respiratory tract infections, lymphoproliferation and hypogammaglobulinemia. APDS is caused by activating heterozygous mutations in *PIK3CD* (APDS1) or in *PIK3R1* (APDS2) [11,12].

Rett syndrome (estimated prevalence 1/15,000) is characterized by a rapid regression in language and motor skills (i.e. repetitive, stereotypic hand movements) after six to eighteen months of normal psychomotor development [13].
The Lowe syndrome or Oculocerebrorenal syndrome (estimated prevalence 1 to 9 /1,000,000) is a multisystem disorder characterized by congenital cataract, intellectual disabilities, glaucoma, postnatal growth retardation and renal tubular dysfunction [14].
The Silver-Russell syndrome (prevalence 1-9 /1,000,000) is characterized by growth retardation with antenatal onset, characteristic facies and limb asymmetry [15].
The Bardet-Biedl syndrome (prevalence estimated at 1 to 9 /1,000,000) is a ciliopathy characterized by a combination of clinical signs including obesity, pigmentary retinopathy, post-axial polydactyly, polycystic kidneys [16].

From now on, we will refer to as phenotype any sign or symptom, disease, defects, and so forth, affecting a patient.
In this study, we present the methods that we developed to extract phenotypes associated with rare diseases from clinical texts in Dr Warehouse® (DrWH), the clinical data warehouse of the Necker Children's hospital. Then, we evaluate the scalability of our approach in the context of high throughput phenotyping.

2. Material

The *Necker Enfants Malades Hospital* (Necker Children Hospital) is a pediatric University hospital belonging to the Assistance Publique Hôpitaux de Paris group (400 pediatric beds, 200 adult beds). The Necker hospital is a national reference center for rare and undiagnosed diseases. The hospital hosts the *Imagine Institute*, a research institute focused on genetic diseases. *Imagine institute* has developed since 2015 a document-based open-source clinical data warehouse oriented toward free-text: Dr Warehouse® (DrWH). DrWH includes a full text search engine based on the Oracle® text module, and contains as of August 2017 more than 3.9 million clinical free-text documents for more than 446,000 patients.
In Table 1, we describe the demographic characteristics of patients in DrWH. We used all the clinical narratives ranging from hospitalization to outpatient visits reports available in DrWH to perform this study. The heterogeneity of the records is illustrated in Table 2 with the distribution of these records by hospital departments and type of reports.
A public demonstration version of DrWH is available at the URL: https://imagine-plateforme-bdd.fr/dwh_pubmed/. For privacy reasons, the demo version does not contain any patient data but has been populated with PubMed abstracts.

The Unified Medical Language System® (UMLS [17]) is assembled by integrating 153 source vocabularies (including HPO, OMIM, MeSH, and so forth…). The UMLS Metathesaurus® contains about 3.2 million concepts, i.e., clusters of synonymous terms coming from various source vocabularies. The UMLS Semantic Network is a much smaller network of 133 semantic types organized in a tree structure. Each Metathesaurus concept is assigned at least one semantic type. The UMLS integrates mostly terms in English (70%), but other language such as French have a non-negligible coverage (3.09% for 397,203 terms).
Orphanet is an online resource gathering and integrating knowledge on rare diseases. Orphanet was established in France in 1997 and became a European initiative now involving a consortium of 40 countries in Europe and the rest of the world. Orphanet data are organized using ontologies and structured data [18]. Orphadata is a partial extraction of the data stored in Orphanet freely accessible and organized as XML files [19].

Table 1: Description of the population of the data warehouse at Necker hospital.

	DrWH
Nb patients	446,481
Sex ratio (M)	47%
Median Nb reports excluding biological reports per patient	2 [1-6]
Median follow up (years) per patient	0.06 [0-2]

In brackets lower and upper quartile.

Table 2: Number of documents per Hospital department and per type of records

Hospital departments	# Documents	Types of records	# Documents
Gyneco-Obstetrics	433,698	Laboratory	1,563,450
Pediatric Cardiology	253,474	Consultation	834,619
Adult Clinical Hematology	227,520	Imaging	379,538
Metabolism-Pediatric Neurology	207,804	Discharge letter	293,342
Nephrology Transplantations Adult	187,388	Diagnostic Related Group	255,312
Pediatric Nephrology	175,041	Hospitalization	226,723
Pediatric Immuno-Hematology	152,226	surgery	111,598
Pediatric Radiology	151,811	Day hospital	88,244
Adult Radiology	150,612	Emergency	41,515
Pediatric Cardiac Surgery	136,272	Exams	31,042
Pediatric Visceral Surgery	121,758	Prescription	24,859
Pediatric Orthopedic Surgery	120,287	Medical certificate	24,222
Adult Nephrology	116,602	Pathology report	24,215
Anesthesia intensive care unit Adult And Pediatric	114,773	Foetopathology	8,858
Pediatric Gastroenterology	113,857	Multidisciplinary consultation meeting	6,605
Emergency	108,367	Other	5,786
General Pediatrics	97,831	Staff meeting reports	3,669
Physiology	88,981	**Total**	**3,923,597**
Pediatric ear nose and throat	82,717		
Pediatric Intensive Care Unit	77,599		
Other	804,979		
Total	**3,923,597**		

3. Method

In this study, we aim at using automated methods to mine the large body of text documents available in the CDW. This section describes the free-text document processing, and details the exploration of phenotypes associated with six use cases.

Processing text-documents.

In a nutshell, we leveraged the UMLS to extract biomedical concepts from patients' text reports. We selected all French terms (including synonyms) from the UMLS Metathesaurus (version 2017AA) and filtered out concepts having less than three characters, or more than 80 characters. We considered only the concepts assigned to one of the

following semantic types: 'Sign or Symptom', 'Disease or Syndrome', 'Finding', 'Pathologic Function', 'Congenital Abnormality', 'Physiologic Function', 'Anatomical Abnormality', 'Neoplastic Process', 'Acquired Abnormality' and 'Mental or Behavioral Dysfunction'.

In the context of rare and undiagnosed diseases, clinical narratives are likely to contain many sentences expressing the absence of phenotypes or describing the family history of the patient. Therefore, detecting negation and family history context was essential. We used an algorithm similar to Context [20], and adapted to French [21], to determine if a concept was present as an affirmative expression or a negative one, and to capture the context (family history or patient information) [22]. Extracted concepts along with the notion of negation and the context are stored in the CDW database. In this study, we considered exclusively the *not negated* concepts associated with the patients (i.e. not associated with their family).

Use cases: exploring phenotypes of rare disease patients

We created six groups of patients each associated with a specific disease. We queried DrWH at Necker hospital using *Rett Syndrome (and not atypical Rett syndrome), Lowe, Silver Russell, Bardet Biedl, DOCK8 deficiency and APDS* as search criteria. We obtained six sets of patients and their associated corpora of clinical documents (*RETT set, LOWE set, SILVER RUSSELL set, BARDET BIEDL set, DOCK8 deficiency set, and APDS set)*. For each patient set, we extracted all the phenotypes corresponding to UMLS concepts as detailed in the previous section (see Figure 1).

To rank the extracted UMLS concepts in terms of relevance, we used two classic metrics (the frequency and the "term frequency–inverse document frequency" - TF-IDF) for each set of patients. For example, our method identified 1022 distinct concepts in the "RETT syndrome" set.

Computing Frequency and TF-IDF

We computed two simple metrics:

- The *frequency*: the frequency of the concept of interest in the results. For example, the frequency of the concept *stereotypy* in the "Rett syndrome" set is 150 (number of patients having at least one mention of *stereotypy* in at least one document) / 209 (number of patients in the set) = 71.8%.
- The *TF-IDF* (term frequency – inverse document frequency) is intended to reflect how important a word is to a document in the data set. For example, the TF-IDF of the concept *stereotypy* in the "Rett syndrome" result set is 0.081 and is computed as follows:

$$TF - IDF\ (c) = \frac{N_c}{N_{tot}} \times \log\left(\frac{P_{tot}}{P_c}\right)$$

N_c: Number of times this concept c is used in the set

N_{tot}: Number of not distinct concepts in the set

P_{tot}: Total number of patients in the DWH with concepts extracted

P_c: Number of patients with concept c in the set

$$TF - IDF\ (\text{Stereotypy}) = \frac{649}{18{,}538} \times \log\left(\frac{446{,}481}{2{,}233}\right) = 0.081$$

Evaluation

Manual evaluation

We considered six use cases. For each of them, a domain expert was asked to browse the highest ranked phenotypes (top-50 concepts) found by DrWH and evaluate their relevance with regard to the disease of interest. We presented each expert with two lists of the top phenotypes: (i) the top-50 phenotypes ranked by descending frequency and (ii) the top-50 phenotypes ranked by descending TF-IDF. The experts classified the phenotypes as relevant or not relevant to the disease.

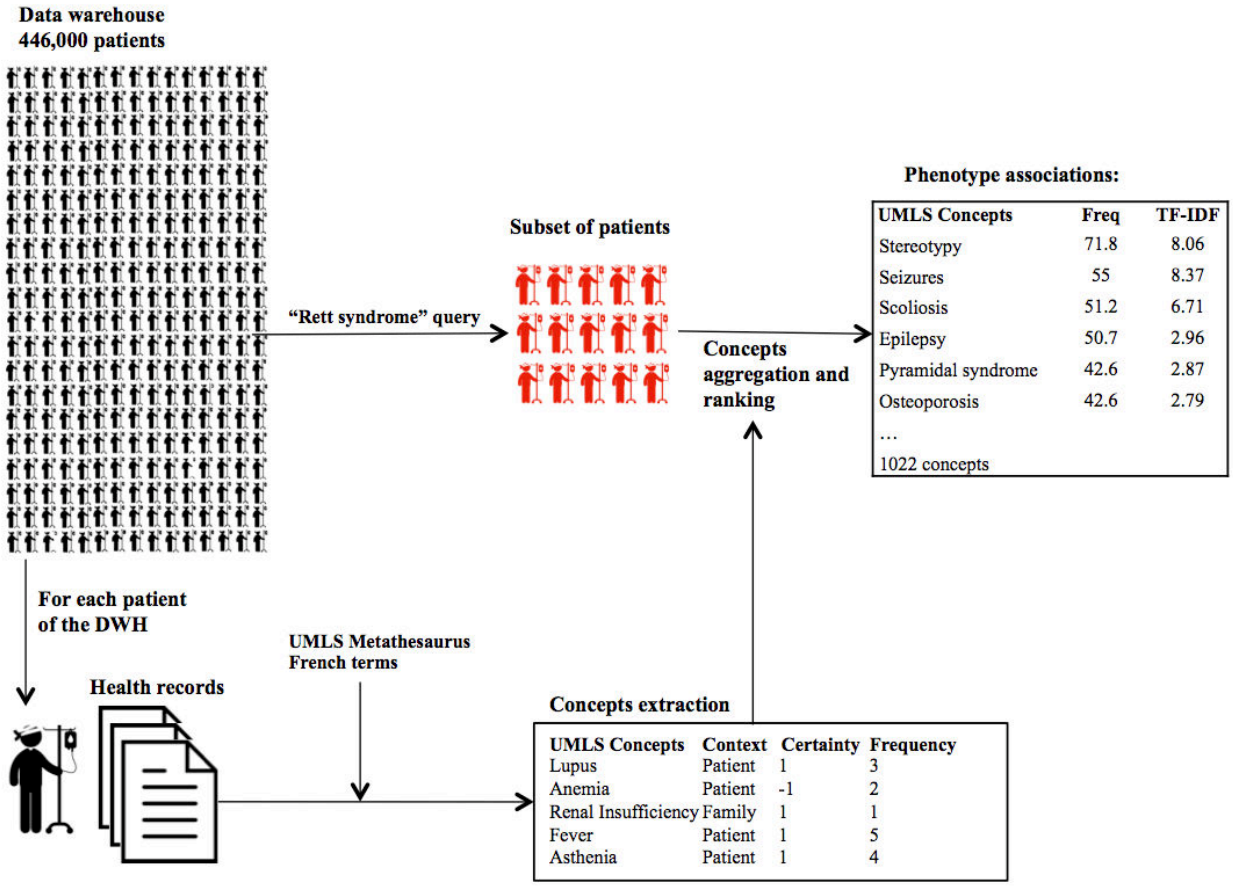


Figure 1: Overview of the method applied to extract concepts from the narrative reports

We stored the number of relevant phenotypes, and their associated ranks. Based on the experts' feedbacks, we computed the Average Precision for each query, and the overall Mean Average Precision. The average precision expresses the correctness of the top ranked. The Mean Average Precision evaluates the average precision across a series of queries [23].

Comparison to Orphadata

For each disease set we compared the UMLS concepts obtained by our method with the concepts in Orphadata with the following steps:

- We extracted the CUIs corresponding to the HPO concepts provided by Orphadata (Orphadata CUI).
- We mapped the CUIs extracted from the corresponding patient set to the Orphadata CUIs. We considered hierarchical links between concepts. Concepts were considered equivalent (i) in case of exact mapping (same CUI) or (ii) when an ascendant of the patient set CUI was successfully mapped to an Orphadata CUI.
- We calculated the number of equivalent concepts and the number of concepts present in only one data source.

For all these steps we limited ourselves to the CUIs with French terms. The steps are illustrated with the example of Rett syndrome in Figure 2.

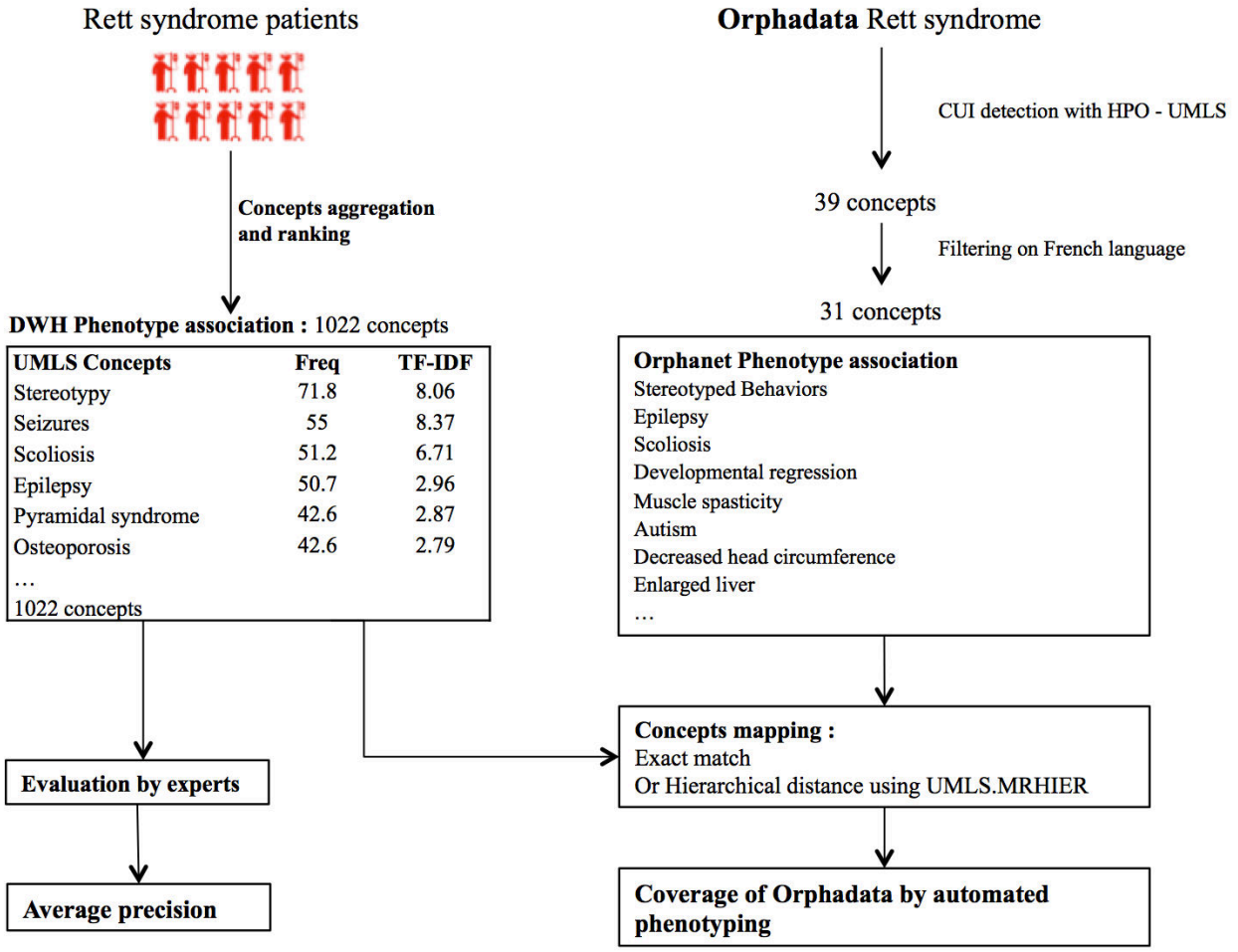


Figure 2: Evaluation procedure for the RETT set

4. Results

Document processing in DrWH

We extracted a total of 18.7 million concepts from 3.9 million medical records, representing 446,481 distinct patients. Among these concepts, 4 % were related to family history. Among the 96% of the remaining concepts, 72% were classified into as not negated expression (12.99 million of concepts) (Table 3).

Table 3: Number of concepts extracted per context and certainty

Context / Certainty	Negated	Not negated
Family history	179,938	522,009
Patient	5,007,517	12,988,474
Total number of concepts	5,187,455	13,510,483

Detailed expert evaluation

The description of the data available in each cohort and the evaluation by the experts are detailed in Table 4. The Figure 3 is a screenshot of the graphical user interface of Dr Warehouse for Rett

The description of the data available in each cohort and the evaluation by the experts are detailed in Table 4. The Figure 3 is a screenshot of the graphical user interface of Dr Warehouse for Rett syndrome. The automated phenotyping identified a vast number of UMLS concepts associated with the diseases. More precisely we identified an average of 768 UMLS concepts associated to each disease. In contrast, the number of UMLS concepts found in Orphadata ranges from 16 for the Silver-Russell syndrome to 120 for the Lowe syndrome. APDS was not documented in Orphanet at the time of redaction of this article. Overall, the experts classified between 11 (SILVER-RUSSELL set, ranked by TF-IDF) and 41 (LOWE set, ranked by TF-IDF) of the top-50 results as relevant to the disease. The number of phenotypes identified by the union of results obtained through ranking by frequency and ranking by TF-IDF ranges from 16 (SILVER-RUSSELL set) to 52 (DOCK8 deficiency and APDS sets).

The Mean Average Precision is 0.79 for results ranked by Frequency and 0.79 for results ranked by TF-IDF.

Table 4: Description and evaluation of the 6 sets of patients

Sets	RETT	DOCK8 deficiency	LOWE	SILVER RUSSELL	BARDET BIEDL	APDS 1 and 2
Median age at visit (years)	8.2 [4.8-12.6]	11.4 [9.3-14.1]	12.8 [5.8-20.3]	2.4 [0.8-5.4]	15.7 [10.1-41.5]	12.8 [7.7-18.6]
Median follow up (years)	2.6 [0-4.9]	3.1 [0.3-9]	6.6 [3-10.3]	2 [0.8-4.7]	2 [0.1-6.6]	7.5 [4.8-8.6]
# Patients	209	15	23	50	53	23
# Documents	5,034	3,296	1,325	1,133	1,317	2,337
UMLS concepts extracted, not negated and in patient context						
# UMLS names	18,538	6,886	5,281	6,563	6,345	9,716
# distinct UMLS concepts	1,022	706	577	738	801	710
Evaluation by experts in the Top50 UMLS concepts						
Medical Experts	NBB	CP	RS	JA	RS	NM
# UMLS concepts ranked by Freq	31	36	36	16	17	39
# UMLS concepts ranked by TF-IDF	38	37	41	11	12	37
# UMLS concepts Freq ∪ TF IDF	42	52	50	16	19	52
# UMLS concepts Freq ∩TF IDF	28	22	28	11	11	25
Average Precision, ranked by Freq	0.86	0.91	0.88	0.55	0.66	0.83
Average Precision, ranked by TF-IDF	0.91	0.95	0.89	0.66	0.52	0.83

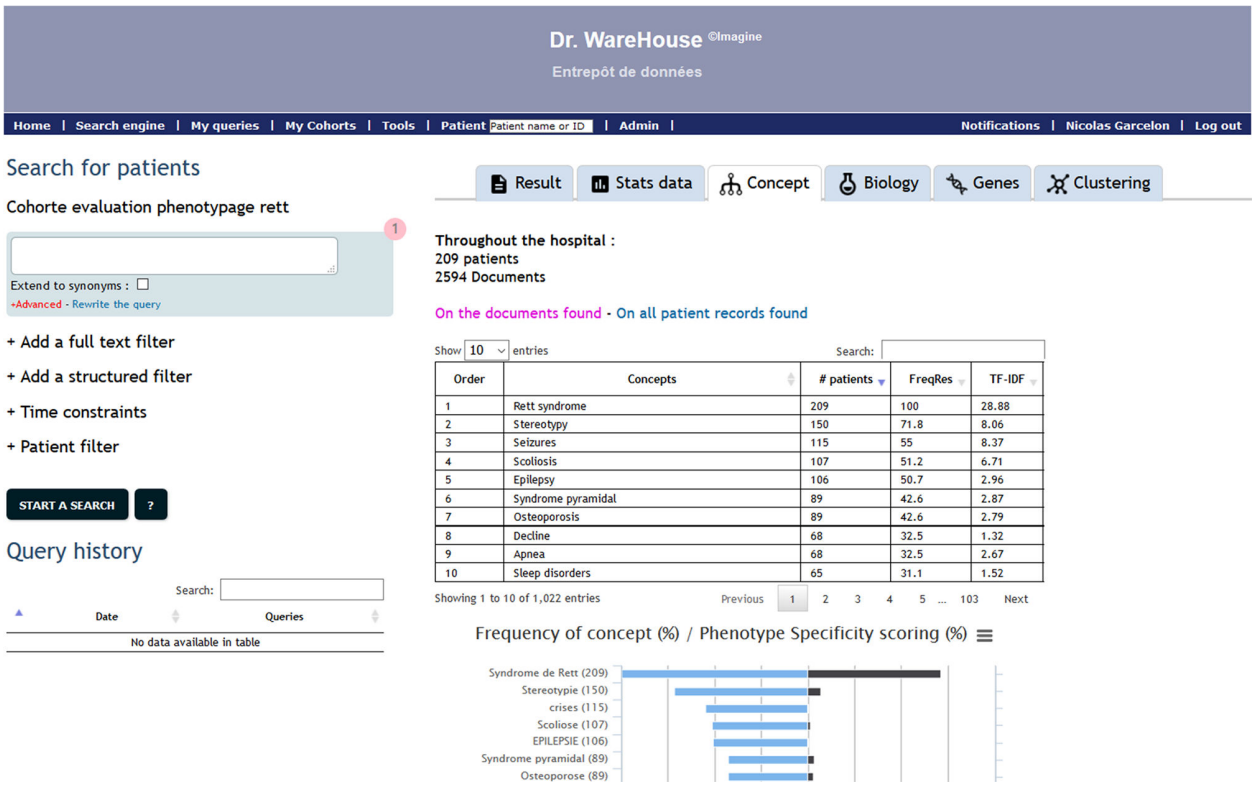


Figure 3: Screenshot of Dr Warehouse and the concept tab for "Rett syndrome" query.

Comparison with Orphadata

The comparison with Orphadata is detailed in Table 5. The limitation to French terms resulted in a reduction of an average of 16 concepts, corresponding to an average of 39% of the CUIs (max: 63%, min 21%).

We obtained the best coverage for the SILVER RUSSEL set with 100% of the Orphadata concepts present in the concepts of the patient set. The lowest coverage was found for the LOWE set with 66% of the 76 Orphadata concepts present in the concepts of the patient set. The average coverage for all the patient sets was 78%.

In the six diseases studied, 2.8% of Orphadata CUIs do not belong to the semantic types used for the automated phenotyping. For example, "Dislocated hips" (HP:0002827) is part of the description of Lowe syndrome in Orphadata and is assigned to the semantic type "Injury or Poisoning" in the UMLS.

Among the CUIs in Orphadata, 41 concepts are not represented in the patient sets CUIs.

5. Discussion

Findings and practical significance

Expert evaluation

The number of documents from the evaluated sets shows a large heterogeneity (average 2,407 +/- 1,528). There are almost five times more documents in the RETT set (5,034) than in the SILVER-RUSSELL set (1,133). The variation of the number of distinct concepts is less important (average 768 +/- 167): the RETT set (1,074) has twice as many concepts as the LOWE set (577).

Table 5: Comparison with Orphadata

	RETT	DOCK8	LOWE	SILVER RUSSELL	BARDET BIEDL	APDS
# Concepts HPO Orphadata (English)	39	18	120	16	25	-
# Concepts HPO Orphadata (French) [A]	31	10	76	6	17	-
# UMLS distinct concepts extracted [B]	1,022	706	577	738	801	710
# [A] ∩ [B] (coverage)	22	7	50	6	14	-
% [A] ∩ [B] / [A] (coverage %)	0.71	0.70	0.66	1.00	0.82	-

The performances are heterogeneous according to the sets evaluated.

The phenotyping proposed by DrWH on RETT, DOCK8 deficiency, LOWE and APDS sets are well evaluated regardless of the measure used to rank the phenotypes. Conversely, DrWH performed not as well on the SILVER-RUSSELL and BARDET-BIEDL sets. However, the average precision (reflecting the ranking of well classified candidates, the higher the average precision is, the highest the best ranked candidates are) is still reasonably high despite a low number of correct phenotypes. The average precision for results ordered by frequency ranges from 0.55 for SILVER-RUSSELL to 0.91 for DOCK8 deficiency, and between 0.52 (BARDET-BIEDL) to 0.95 (DOCK8 deficiency) for results ordered by TF-IDF.

Enrichment of disease description in Knowledge Bases

Our approach can be used to enrich existing phenotypic description of rare diseases. For example, osteoporosis was significantly associated with Rett syndrome in the Necker data warehouse. The association is present neither in Orphanet nor in OMIM. It is however described in six articles in Medline [24–29]. This example illustrates the advantages of using the UMLS to integrate several data sources (the UMLS demonstrated its ability to support the linkage between the terminologies used in the different sources). Using the UMLS as a pivot terminology, it is possible to explore phenotypes gathered in clinical data warehouse, knowledge bases and the literature.

Moreover, this method enables a quick exploration of phenotypes in a population. This feature is especially meaningful in the case of rare diseases for which the information is scarce. In a research context, we have shown with the six examples, that our method is able to automatically display the phenotypes associated with rare diseases in a cohort of patients. The same approach could be used to look for undescribed phenotypes associated with new mutations (using gene names as a query for example or a series of patients selected manually). The Necker hospital and *Imagine* Institute collaborate actively to increase the knowledge on rare diseases and the concept explorer from the CDW is used on a daily basis by the staff to support translational research: When a geneticist discovers a new mutation, the exploration of the documents gathered from patients presenting the mutation in the CDW can support the description of the associated phenotypes. For example, the phenotypes associated to APDS 1 and 2 could provide basis for the description of the syndrome.

DrWH can be used to assist experts in the identification of phenotypes of interest. After careful review and comparison with other cohorts, such associations could be used to enrich online reference resources. Moreover, the method is easily reproducible, and the comparison of phenotypes coming from a variety of clinical data

warehouses can provide candidates (union of the candidate phenotypes) or reinforce the interest on specific candidate phenotypes (using the intersection of different submissions).

In addition, the prevalence of signs and symptoms for a given disorder can be estimated using the frequencies provided by DrWH. Our method can provide the clinicians with prevalence of phenotypes in addition to the associations. In our running example Rett syndrome, "stereotypy" had a prevalence of 71.8%, consistent with Orphanet (Very frequent 80-99%); similarly "scoliosis" had a prevalence of 51.2%, vs. frequent (30-79%) in Orphanet. Conversely, the prevalence of "apraxia" in DrWH was 12.9%, whereas apraxia is considered very frequent (80-99%) in the Rett syndrome by Orphanet. A more precise estimation of the frequency would require considering not only single concepts but also group of semantically closed concepts.

Limitations

Comparison to a gold standard and, interoperability issues

The automated evaluation of phenotypes found by DrWH by comparison to a gold standard (e.g. Orphanet with Orphadata) is complex.

(1) We leverage the French terms from the UMLS. The coverage of French term is limited compared to the extent of the English counterpart. For example, in Orphadata the Rett syndrome is associated with 39 phenotypes, of which 31 exist in French in the UMLS (Table 5). The difference is more dramatic with the Lowe syndrome: for 120 phenotypes, only 76 have a French counterpart. Our automated exploration is based on the use of medical terminologies in French, and DrWH cannot recognize a phenotype that is not present in French. For example, Triangular Face (HPO: HP:0000325, UMLS: C1835884) is a sign associated to Silver-Russell syndrome in Orphadata and is absent from the UMLS concepts extracted from the corresponding set. Nevertheless, 27 patients of the SILVER-RUSSELL syndrome set have "face triangulaire" in their narrative records, but the concept "Triangular Face" does not exist in French in the UMLS. Note that the current version of DrWH enables nonetheless relevant explorations, and allows the discovery of phenotypes of interest.

(2) Different granularity between concepts extracted from Orphadata and DrWH can occur and cannot be addressed by simple hierarchical reasoning. Concepts may be related semantically, but not identical nor hierarchically linked (e.g. *Hypotonia (CUI0026827) vs muscle weakness (CUI 151786)*).

(3) Some concepts are not in the semantic types that we used for the automated phenotyping. In Orphadata, "autoagression" is a sign associated to Rett syndrome. We found 10 patients in the RETT set with "automutilation" in their narrative reports, but this concept is in the semantic type "Injury or Poisoning".

Study population

Our warehouse hosts data produced by a children hospital, and therefore, phenotypes can be different from adult patients (for example, Alzheimer disease is not represented in pediatrics). However, patients with rare diseases may be followed-up at the Necker hospital even during adulthood, enabling long longitudinal data collection. Longitudinal follow-up makes possible to observe the age of apparition of the phenotypes and reconstruct the natural history of rare diseases.

Related work

Information extraction

Several approaches have been developed to extract information from free-text records. Savova *et al.* [30] developed cTAKES, an open source modular system of pipelined components combining rule-based and machine learning techniques. cTAKES

aims at the extraction of information from the clinical narratives. Despite development in other languages [31,32], most of the open source clinical Natural Language Processing systems have been developed for the English language (MedLee [33], MetaMap [34], HITex [35]). Approaches have been developed specifically for the French language. More recently a challenge was dedicated to the extraction of information in multiple language medical documents (including French)[36].

Phenotype associations

Denny *et al*. [37] have developed the PheWAS (Phenome Wide Association study) method to mine associations between a wide range of phenotypes and a genotype to explore. Their PheWAS was based on ICD9 billing codes. Building on this work, Neuraz *et al*. [38] used data extracted from free-text reports to restrict the phenotypes used in the analysis to a delimited period of time. Hebbring *et al*. [39] augmented Denny's work by developing a PheWAS study using information extracted from clinical text. They concluded on the relevance of text-based approaches compared to the ICD9-based approach.

Narrative reports versus coded data

We have shown that text exploration of clinical reports can provide phenotypes of interest. Whereas structured databases are particularly adapted for the collection of data regarding well documented diseases, clinical report based exploration enables the secondary use of data collected during care. Such approaches allow the development of learning health systems in which there is a bidirectional relation between routine care data and research. In addition, patient generated data could be integrated and mined along with the EHRs [40].

We plan to conduct additional studies by comparing our results with the French rare disease registry [41].

Scalability

The concept explorer of Dr Warehouse enables the exploration of millions of clinical narratives in a simple manner. The algorithm is optimized to display the phenotype analysis of thousands of documents quickly, and limited expertise is needed to write and execute queries. The queries demonstrated in this study only took a few seconds to run, enabling a real time exploration of the data. The expert user can easily sort the associated phenotypes according to their need, depending on the use case.

6. Conclusion

Clinical Data Warehouses can be used to perform Next Generation Phenotyping, especially for rare diseases. We have developed a method to detect phenotypes associated with a group of patients using medical concepts extracted from free-text clinical narratives. There are still hurdles to overcome with terminologies other than in English, however experts' evaluation suggests that the phenotypes identified using the Frequency and TF-IDF can be useful to populate knowledge bases in addition to literature mining.

7. Declarations

Ethics approval and consent to participate

We got an ethical approval by the French IRB CPP Il-de-France II (IRB registration number 00001072) registered under reference 2016-06-01.

Consent for publication

Not applicable

Availability of data and material

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

NG, AN and VB made substantial contributions to the acquisition of data. NG, AN, BR and AB conceived the hypothesis and designed the study. RS, JA, NBB, CP and NM manually evaluated the automated phenotyping. All authors made substantial contributions to the analysis and interpretation of data, were involved in drafting and critically revising the manuscript, gave final approval of the version to be published and agree to be accountable for all aspects of the work.

Acknowledgements

BR is supported in part by the SIRIC CARPEM cancer integrated research program.

8. Figures

Figure 1: Overview of the method applied to extract concepts from the narrative reports

Figure 2: Evaluation procedure for the RETT set

Figure 3: Screenshot of Dr Warehouse and the concept tab for “Rett syndrome” query.

9. References

1. Office of the National Coordinator for Health Information Technology. Health Record Adoption: 2004-2014, Health IT Quick-Stat #50. [Internet]. 2015 Sep. Available from: dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php.

2. Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, et al. Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. Health Aff. Proj. Hope. 2015;34:2174–80.

3. Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud. Health Technol. Inform. 2010;160:193–7.

4. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J. Am. Med. Inform. Assoc. JAMIA. 2010;17:124–30.

5. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: the Vanderbilt approach. J. Biomed. Inform. 2014;52:28–35.

6. Raghavan P, Chen JL, Fosler-Lussier E, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci. 2014;2014:218–23.

7. Escudié J-B, Jannot A-S, Zapletal E, Cohen S, Malamut G, Burgun A, et al. Reviewing 741 patients records in two hours with FASTVISU. AMIA. Annu. Symp. Proc. 2015;2015:553–9.

8. Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. J. Am. Med. Inform. Assoc. JAMIA. 2015;22:76–85.

9. Radico - Rare Disease Cohorts [Internet]. [cited 2017 Sep 30]. Available from: http://www.radico.fr/en/accueil

10. RESERVED IU--AR. Orphanet: Combined immunodeficiency due to DOCK8 deficiency [Internet]. [cited 2017 Sep 30]. Available from: http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=EN&Expert=217390

11. Picard C, Al-Herz W, Bousfiha A, Casanova J-L, Chatila T, Conley ME, et al. Primary Immunodeficiency Diseases: an Update on the Classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency 2015. J. Clin. Immunol. 2015;35:696–726.

12. RESERVED IU--AR. Orphanet: Activated PI3K delta syndrome [Internet]. [cited 2017 Sep 30]. Available from:

http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=EN&Expert=397596
13. RESERVED IU--AR. Orphanet: Rett syndrome [Internet]. [cited 2017 Sep 30]. Available from: http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=EN&Expert=778
14. RESERVED IU--AR. Orphanet: Oculocerebrorenal syndrome of Lowe [Internet]. [cited 2017 Sep 30]. Available from: http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=EN&Expert=534
15. RESERVED IU--AR. Orphanet: Silver Russell syndrome [Internet]. [cited 2017 Sep 30]. Available from: http://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=EN&Expert=813
16. RESERVED IU--AR. Orphanet: Bardet Biedl syndrome [Internet]. [cited 2017 Sep 30]. Available from: http://www.orpha.net/consor/cgi-bin/OC_Exp.php?Expert=110
17. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf. Med. 1993;32:281–91.
18. Orphanet: an online rare disease and orphan drug data base. Copyright, INSERM 1997. [Internet]. [cited 2017 Sep 22]. Available from: http://www.orpha.net.
19. INSERM. Orphadata: Free access data from Orphanet. © INSERM 1997. Available on http://www.orphadata.org. Data version (XML data version) [Internet]. 1997 [cited 2017 Sep 24]. Available from: http://www.orphadata.org/cgi-bin/inc/product4.inc.php
20. Harkema H, Dowling JN, Thornblade T, Chapman WW. Context: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. J. Biomed. Inform. 2009;42:839–51.
21. Chapman WW, Hillert D, Velupillai S, Kvist M, Skeppstedt M, Chapman BE, et al. Extending the NegEx lexicon for multiple languages. Stud. Health Technol. Inform. 2013;192:677–81.
22. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. J Am Med Inform Assoc.
23. Beitzel SM, Jensen EC, Frieder O. MAP. In: LIU L, ÖZSU MT, editors. Encycl. Database Syst. [Internet]. Springer US; 2009 [cited 2017 Sep 30]. p. 1691–2. Available from: http://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_492
24. Bahi-Buisson N. Genetically determined encephalopathy: Rett syndrome. Handb. Clin. Neurol. 2013;111:281–6.
25. Budden SS, Gunness ME. Possible mechanisms of osteopenia in Rett syndrome: bone histomorphometric studies. J. Child Neurol. 2003;18:698–702.
26. Cortelazzo A, De Felice C, Guerranti R, Signorini C, Leoncini S, Pecorelli A, et al. Subclinical Inflammatory Status in Rett Syndrome. Mediators Inflamm. [Internet]. 2014 [cited 2017 Sep 30];2014. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3913335/
27. Jefferson A, Leonard H, Siafarikas A, Woodhead H, Fyfe S, Ward LM, et al. Clinical Guidelines for Management of Bone Health in Rett Syndrome Based on Expert Consensus and Available Evidence. PLoS ONE [Internet]. 2016 [cited 2017 Sep 30];11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4743907/
28. Lotan M, Reves-Siesel R, Eliav-Shalev RS, Merrick J. Osteoporosis in Rett syndrome: a case study presenting a novel management intervention for severe osteoporosis. Osteoporos. Int. J. Establ. Result Coop. Eur. Found. Osteoporos. Natl. Osteoporos. Found. USA. 2013;24:3059–63.
29. Zysman L, Lotan M, Ben-Zeev B. Osteoporosis in Rett syndrome: A study on normal values. ScientificWorldJournal. 2006;6:1619–30.
30. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System

(cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. JAMIA. 2010;17:507–13.
31. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. PLOS Comput Biol. 2011;7:e1002141.
32. Deléger L, Grouin C, Zweigenbaum P. Extracting medication information from French clinical texts. Stud. Health Technol. Inform. 2010;160:949–53.
33. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J. Am. Med. Inform. Assoc. JAMIA. 2004;11:392–402.
34. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc. AMIA Annu. Symp. AMIA Symp. 2001;17–21.
35. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med. Inform. Decis. Mak. 2006;6:30.
36. CLEF e-health 2016 [Internet]. Available from: https://sites.google.com/site/clefehealth2016/task-2
37. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinforma. Oxf. Engl. 2010;26:1205–10.
38. Neuraz A, Chouchana L, Malamut G, Le Beller C, Roche D, Beaune P, et al. Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. PLoS Comput. Biol. 2013;9:e1003405.
39. Hebbring SJ, Rastegar-Mojarad M, Ye Z, Mayer J, Jacobson C, Lin S. Application of clinical text data for phenome-wide association studies (PheWASs). Bioinforma. Oxf. Engl. 2015;
40. Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J. Am. Med. Inform. Assoc. JAMIA. 2015;22:43–50.
41. Maaroufi M, Choquet R, Landais P, Jaulent M-C. Towards data integration automation for the French rare disease registry. AMIA Annu. Symp. Proc. AMIA Symp. 2015;2015:880–5.

3.4. Similarité entre patients

Devant prendre en charge une patiente avec un tableau clinique particulièrement complexe et la littérature faisant défaut pour prendre une décision clinique, Frankovich relate dans un article publié dans le New England qu'ils ont utilisé leur entrepôt de données pour retrouver des patients similaires à celle-ci. Ils ont ainsi en quelques heures pu analyser les dossiers de ces patients et choisir le traitement permettant la meilleure prise en charge (Frankovich et al., 2011).

Si un moteur de recherche a pu suffire pour retrouver les patients comparables dans le retour d'expérience de Frankovich, il peut s'avérer compliquer d'exprimer la liste nécessaire et suffisante de signes permettant de retrouver des patients similaires à un patient index. En effet une recherche sur des critères Booléens peut être soit trop spécifique soit trop sensible.

L'institut *Imagine* diagnostique des patients avec de nouvelles mutations non décrites dans la littérature. L'objectif est de pouvoir retrouver parmi les patients de l'hôpital Necker des patients similaires au patient diagnostiqué et éligibles pour un test génétique sur le même gène. Il s'agit alors de proposer une distance de similarité permettant de ne pas focaliser ni sur l'ensemble des signes cliniques du patient index ni sur l'ensemble des signes des patients de Necker mais de proposer des patients qui ont une proximité phénotypique avec le patient index. Nous présentons ici une méthode utilisant le Vector Space Model (Salton, 1968) sur les concepts extraits depuis les comptes rendus médicaux des patients.

Nous avons évalué notre méthode sur 233 patients index répartis dans cinq cohortes : Syndrome de PI3K-delta activé, Syndrome de Lowe, Epidermolyse bulleuse dystrophique, Epidermolyse bulleuse de type Dowling Meara, Syndrome de Rett. Pour chaque patient index testé, nous avons évalué la capacité de notre méthode à retrouver les patients de la cohorte du patient index parmi les 400 000 patients de l'entrepôt. Nous avons préalablement supprimé les concepts contenant le syndrome pour les patients de la cohorte. La similarité a ainsi été calculée uniquement sur les phénotypes associés.

Nous avons déterminé qu'il faut un minimum de 8 concepts en commun (considéré comme non négatif) avec le patient index pour optimiser le calcul et réduire le nombre de patients sur lequel réaliser la similarité.

Nous avons comparé le calcul de similarité avec et sans une pondération des concepts par le TF-IDF. Et nous avons évalué la sensibilité et la spécificité avec ou sans la prise en compte des concepts classés comme négatifs. Nous montrons que la pondération par le TF-IDF améliore incontestablement le rappel tandis que la prise en compte des concepts négatifs ne l'améliore que légèrement.

En moyenne, un patient index renvoie 51% de patients vrais positifs dans le top 30 des patients les plus similaires, c'est à dire qu'ils appartiennent à la cohorte du patient index. Si le rappel reste globalement relativement bas (0.20 en moyenne), nous considérons notre méthode comme

pertinente dans le contexte des maladies rares ou non diagnostiquées. En effet, si un seul patient permet de diagnostiquer ne serait-ce qu'un patient non diagnostiqué dans l'entrepôt de données, nous estimons cela comme une victoire sur l'errance diagnostique des patients. Nous discutons plus en détail ces résultats dans l'article publié dans le Journal of Biomedical Informatics en août 2017.

Journal of Biomedical Informatics 73 (2017) 51–61

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack

Nicolas Garcelon [a,b,c,*], Antoine Neuraz [c,d], Vincent Benoit [a,b], Rémi Salomon [a,b,e], Sven Kracker [a,b,f], Felipe Suarez [a,b,g], Nadia Bahi-Buisson [a,b,h], Smail Hadj-Rabia [a,b,i], Alain Fischer [a,b,j,k,l], Arnold Munnich [a,b,m,n], Anita Burgun [c,d,o]

[a] Institut Imagine, Paris Descartes Université Paris Descartes-Sorbonne Paris Cité, Paris, France
[b] INSERM, Institut Imagine, UMR 1163, Université Paris Descartes, Sorbonne Paris Cité, Paris, France
[c] INSERM, Centre de Recherche des Cordeliers, UMR 1138 Equipe 22, Université Paris Descartes, Sorbonne Paris Cité, Paris, France
[d] Département d'informatique médicale, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[e] Service de Néphrologie Pédiatrique, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[f] Laboratory of Human Lymphohematopoiesis, INSERM UMR 1163, Paris, France
[g] Service de Hématologie, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[h] Service de neurologie pédiatrique, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[i] Service de Dermatologie, Centre de Références maladies Génétiques à Expression Cutanée (MAGEC), Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[j] Centre de Référence Déficits Immunitaires Héréditaires, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[k] Unité d'Immunologie-Hématologie et Rhumatologie Pédiatrique, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[l] College de France, Paris, France
[m] Département de génétique médicale, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France
[n] Centre de Référence des Maladies Osseuses Constitutionnelles, INSERM UMR 1163, Laboratoire de bases moléculaires et physiopathologiques de l'ostéochondrodysplasie, Paris Descartes-Sorbonne Paris Cité University, AP-HP, Institut Imagine, 75015 Paris, France
[o] Hôpital Européen Georges Pompidou, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France

ARTICLE INFO

Article history:
Received 24 February 2017
Revised 5 July 2017
Accepted 24 July 2017
Available online 25 July 2017

Keywords:
Data warehouse
Vector space model
Similarity measures
Rare diseases
Electronic health records

ABSTRACT

Objective: In the context of rare diseases, it may be helpful to detect patients with similar medical histories, diagnoses and outcomes from a large number of cases with automated methods. To reduce the time to find new cases, we developed a method to find similar patients given an index case leveraging data from the electronic health records.

Materials and methods: We used the clinical data warehouse of a children academic hospital in Paris, France (Necker-Enfants Malades), containing about 400,000 patients. Our model was based on a vector space model (VSM) to compute the similarity distance between an index patient and all the patients of the data warehouse. The dimensions of the VSM were built upon Unified Medical Language System concepts extracted from clinical narratives stored in the clinical data warehouse. The VSM was enhanced using three parameters: a pertinence score (TF-IDF of the concepts), the polarity of the concept (negated/not negated) and the minimum number of concepts in common. We evaluated this model by displaying the most similar patients for five different rare diseases: Lowe Syndrome (LOWE), Dystrophic Epidermolysis Bullosa (DEB), Activated PI3K delta Syndrome (APDS), Rett Syndrome (RETT) and Dowling Meara (EBS-DM), from the clinical data warehouse representing 18, 103, 21, 84 and 7 patients respectively.

* Corresponding author at: Imagine – Institute for genetic diseases, 24 boulevard du Montparnasse, 75015 Paris, France.
E-mail address: nicolas.garcelon@institutimagine.org (N. Garcelon).

http://dx.doi.org/10.1016/j.jbi.2017.07.016
1532-0464/

Results: The percentages of index patients returning at least one true positive similar patient in the Top30 similar patients were 94% for *LOWE*, 97% for *DEB*, 86% for *APDS*, 71% for *EBS-DM* and 99% for *RETT*. The mean number of patients with the exact same genetic diseases among the 30 returned patients was 51%.
Conclusion: This tool offers new perspectives in a translational context to identify patients for genetic research. Moreover, when new molecular bases are discovered, our strategy will help to identify additional eligible patients for genetic screening.

1. Introduction

There are approximately 7000 different types of rare diseases and disorders affecting a large population worldwide [1]. Collecting and analyzing patient data is critical aspect of rare and undiagnosed diseases programs that has the potential to increase the opportunity of discovering new knowledge. Because of their individual rarity, identifying patients that share similar phenotypes can be particularly challenging. To quote E. Ashley, the co-chair of the steering committee for the Undiagnosed Diseases Program, the challenge with rare disease data "is not so much finding the needle in the haystack as finding the right needle in a whole pile of needles" [2]. It may be helpful to detect patients with similar medical histories, diagnoses and outcomes from a large number of cases with automated methods. For example, when a new causal mutation for a disease is discovered, being able to find potential cases in a retrospective database by looking for patients similar to the 2 or 3 patients already diagnosed and genotyped could reduce their diagnosing journey and confirm the causal association.

The ability to retrospectively mine all patient records has proven beneficial [3]; and it has been made possible with the widespread adoption of clinical data warehouses such as i2b2 [4] and STRIDE [5] in hospitals. In most cases, the users query the data warehouse using a Boolean combination of criteria e.g., a list of signs and phenotypic traits. But in the case of rare complex disorders, it may be difficult to query the data warehouse in the quest for similar patients, with a formal and precise list of symptoms. The search for similar cases in the data warehouse based on similarity metrics would be more powerful and versatile.

We have adapted the vector space approach to provide similarity metrics between rare disease patient medical records. In this paper, we present the method, its implementation and its evaluation on the Necker-*Enfants Malades*/*Imagine* data warehouse.

2. Background and significance

Both supervised and unsupervised machine learning methods [6] have been used to compute patient similarity, e.g., support vector machines. All these methods are based on learning models thus require training sets of a sufficient number of cases.

The Vector space model (VSM) was first proposed by Salton in 1968 [7]. Then it was implemented in SMART [8], an information retrieval system that computes similarity between documents represented as vectors of keywords. The matrix (documents by indexing terms) consists of binary values indicating the presence (1) or absence (0) of a term in a document. Salton demonstrated in 1968 the noticeable improvement in performance by using the term frequency weight [9] instead of binary values. Spärck Jones (1972) demonstrated the interest of the frequency of a term in a collection [10] and introduced the tf-idf (term frequency – inverse document frequency) weight.

Since SMART, the VSMs have been broadly used in information retrieval [11–17], in classification [18,19], and clustering [20]. The VSM has also been used in other applications. For example in social network analysis, Lee et al. used the VSM to find new network ties [21]. Santos et al. used Topic-based VSM approach to enhance a spam filter [22]. Castells et al. also used ontology in association with the VSM to improve the ranking of a search engine [23].

The VSM has also been used to compute similarity in the biomedical domain.

With the objective of identifying potentially related diseases based on genetic relationships, Sarkar et al. [24] proposed an adaptation of the VSM that bridges gene and disease knowledge inferred across three knowledge bases: Online Mendelian Inheritance in Man, GenBank, and Medline. In this study, the relatedness between diseases, via this network of gene-based relationships, was determined using a cosine similarity metric. The authors concluded that VSM was a potentially powerful method for exploring the complex landscape of polygenetic diseases. Lee et al. [25] used a VSM and applied a cosine-similarity-based patient similarity metrics to an intensive care unit database to identify patients who are most similar to each index patient and predict their outcomes. They applied a VSM to MIMIC II structured data, including ICD-9 codes and quantitative data, and showed that their approach outperformed the standard severity scores usually used in the intensive care units.

In a study published in 2013, we applied a VSM approach to identify surgical site infections after neurosurgical procedures in full-text reports [26]. The method applied to patient narrative documents achieved a high recall score (92%) and a precision of 40%, much higher than the same approach based on ICD-10 codes (85% and precision 5%). These results are consistent with several studies that demonstrated that information extraction from unstructured clinical narratives is essential to most clinical applications [27]. For example, structured data alone is insufficient in resolving eligibility criteria for recruiting patients onto clinical studies [28].

All of these studies have suggested that the VSM approach can be effective at representing and computing similarity between patient reports. In the next section, we describe the VSM-based system that we have developed to search for similar patients attending the *Imagine* Institute, a research and healthcare institute focusing on genetic and rare diseases associated with the *Necker-Enfants Malades* Hospital in Paris.

3. Methods

The goal of this study was to explore the potential of using a VSM approach to identify potentially similar patients in a rare disease data warehouse.

3.1. Data warehouse

Necker Enfants Malades Hospital is an Assistance Publique-Hôpitaux de Paris (AP-HP) children's hospital located in Paris, France. The hospital is specialized in rare diseases and is associated with a research institute, the *Imagine* Institute of Genetic Diseases. The hospital and the institute hold a joint clinical data warehouse,

Dr. Warehouse (DrWH) of 3 million electronic health records for about 400,000 patients. In this study, we considered only the narrative reports written by clinicians (i.e. about 2 million narrative reports). We excluded the lab test reports automatically produced in a text format.

As part of the "Extract Transform and Load" (ETL) DrWH pipeline, natural language processing identifies named entities and their context, i.e., negation and subject (patient or family history) [29]. Since the diagnosis of rare diseases is based on an extensive phenotypic description, a detailed characterization of the symptoms, including the negative findings is documented in the patient record. In addition, in the context of genetic diseases, narrative reports usually include a large amount of detailed information associated with patient's relatives. Therefore, it was crucial to distinguish concepts used to describe patient condition versus family history, and negated concepts versus affirmative or dubitative ones [30–32]. The output of this pipeline was a list of concepts mapped to the Unified Medical Language System (UMLS®) Metathesaurus®. Each concept has been classified as (i) belonging to the family history or patient information, and as (ii) negated or not negated information.

We represented patients as vectors of concepts extracted from all their narrative reports, following these rules:

- we excluded the concepts classified in the family history context
- we filtered the concepts based on their UMLS Semantic Types and kept only the following types: 'Sign or Symptom', 'Disease or Syndrome', 'Finding', 'Pathologic Function', 'Congenital Abnormality', 'Physiologic Function', 'Anatomical Abnormality', 'Neoplastic Process', 'Acquired Abnormality'.
- we distinguished between affirmative and negated concepts (e.g., "no adenopathy")
- we concatenated multiple narrative reports in a single record
- we calculated the number of times the concept appeared in the patient record

We then computed similarity metrics, and evaluated the metrics as described below.

3.2. Evaluation set

We considered five rare diseases: Lowe syndrome (*LOWE*), dystrophic epidermolysis bullosa (*DEB*), activated PI3K delta syndrome (*APDS*), Rett syndrome (*RETT*) and Dowling Meara (*EBS-DM*), with prevalence ranging from <1/1 Million for APDS, 1–9/1 Million for Lowe syndrome, 1–9/1 Million for DEB, <1/1 Million in Scotland for EBS-DM, 1–9/100,000 for Rett syndrome [33]. For each disease, the domain experts from the Imagine institute provided a list of genetically diagnosed patients. All these patients had their medical records stored in the Imagine/Necker Enfants Malades data warehouse. As a typographic convention, we refer to this list of genetically diagnosed patients provided by the expert using the term in italic upper case, e.g., *LOWE*, and call these datasets "cohorts".

The rationale for choosing these diseases was as follows:

- Lowe syndrome, dystrophic epidermolysis bullosa, EBS-DM and Rett syndrome are well known diseases with identified genetic causes: mutations in genes *OCRL1* [34], *COL7A1* [35], *KRT5* or *KRT14* [36] and *MECP2* [37] respectively.
- APDS syndrome is a genetic disease with two subtypes: APDS1 is associated with mutation in the *PIK3CD* gene [38] (a catalytic subunit), and APDS2 is associated with mutation in *PIK3R1* gene [39] (a regulatory subunit). These mutations were discovered recently in 2013 and 2014, respectively.

We considered these cohorts provided by the experts as the gold standard for the evaluation of our method. We kept patients with at least 10 concepts (excluding negated concepts) for the analysis, as records with less than 10 concepts correspond mostly to patients that were seen at Necker Hospital for only one consultation.

For each diagnosed index patient, we computed a ranked list of patients from the data warehouse based on their similarity to this index patient.

3.3. Similarity metrics

Considering an index patient (P_{index}) whose record is represented in the VSM as a vector of m concepts, the similarity metrics was computed as follows:

(1) Starting with P_{index}, we selected a set of n patients (P) from DrWH that had at least k concepts in common with P_{index}. k was regarded as a hyper-parameter that could be tailored to specific use cases. This parameter was used to reduce the scope of search to patients with sufficient information and limit the size of P.

(2) We considered $C = \{C_1, C_2, \ldots, C_m\}$, the set of m concepts representing the P_{index}. C was used to define the dimensions of the vector space. For each P_i from P and each concept C_j from C, we computed the TF-IDF weight w_i^j [40]. TF_i^j was the term frequency, i.e., the number of occurrences of C_j for P_i divided by the number of concepts for this patient. IDF^j was the inverse document frequency, i.e., the logarithm of n (number of patients from P) divided by the number of patients with this concept (n_j).

The weight w_i^c of concept c for patient i was calculated as the product of the TF and IDF.

Thus, we obtained the following matrix (1):

$$\begin{pmatrix} \vec{P}_{index} \\ \vdots \\ \vec{P}_n \end{pmatrix} = \begin{pmatrix} w_{index}^1 & \cdots & w_{index}^m \\ \vdots & \ddots & \vdots \\ w_n^1 & \cdots & w_n^m \end{pmatrix} \quad (1)$$

The distance between the index patient (P_{index}) and a patient (P_i) was the cosine of the two vectors (2) representing these patients:

$$\cos(\vec{P}_{index}, \vec{P}_i) = \frac{\vec{P}_{index}\vec{P}_i}{|\vec{P}_{index}||\vec{P}_i|} \quad (2)$$

Subsequently, all P_i were sorted in descending order based on their similarity with P_{index}, and the top 200 (top200) patients were displayed.

3.4. Implementation

The algorithm was implemented using PHP v.5.4 language, Oracle® 11 g for processing, HTML 4 and jQuery 1.11.

4. Evaluation

The aim was to evaluate the ability to find other cases (considered as new patients) by using a similarity distance with an index case (diagnosed patient). The purpose of our method is to find new patients having the same clinical phenotypes (signs or symptoms) of an index patient. To evaluate the ability to find these patients only based on their clinical data; we removed the concept corresponding to the diagnosis of the index patient.

We simulated new patients using two scenarios:

1. We computed the similarity metrics using the vectors with all the concepts but the diagnosis, e.g., for patients with a Lowe syndrome, we removed the concept "Lowe syndrome". Such method is appropriate to evaluate our approach in situations where the diagnosis is made early after birth.
2. We computed the similarity metrics using only the vectors of concepts before the diagnosis, e.g. we removed all the concepts occurring after the date of the diagnosis. Such method is required to avoid biases that could be generated by the diagnosis and is more appropriate when time to diagnosis is long.

4.1. First evaluation: removing only the diagnosis concepts

In order to evaluate our method, we selected a cohort of patients diagnosed with a specific rare disease *D* from the data warehouse. The concept *D* was removed from their bunch of concepts to compute similarity metrics only on signs and symptoms. We repeated this on the five different cohorts with five specific diseases described in the evaluation set section.

Let us consider *LOWE* patients. For each patient in the data warehouse, we removed the corresponding diagnosis concepts (i.e., "Lowe syndrome" and synonym "Oculocerebrorenal syndrome", UMLS CUI C0028860) from his/her list of concepts. Since we limited the information extraction to specific semantic types, all the concepts belonging to the UMLS semantic type 'Gene or Genome' were excluded, such as the gene associated with Lowe syndrome, *OCRL1*.

Then, we repeated the following steps for all the patients in *LOWE*:

(1) We picked one patient (Index Patient) from *LOWE*, the others were mixed with the 400,000 patients of the data warehouse
(2) We calculated the similarity distance between each patient of the data warehouse and the Index Patient with at least *k* concepts in common. To evaluate the similarity distance we varied three parameters:
   a. k concepts in common varying from k = 0 to k = 15 (excluding negated concepts)
   b. The weight parameter used to calculate the distance similarity, with two values: (1) not weighted and (2) weight = TF-IDF
   c. The exclusion of negated concepts (certainty = −1) or their inclusion, i.e., we considered only assertive concepts (certainty = 1) versus all the concepts
(3) We identified the S most similar patients (topS) to the index patient
(4) We computed the recall, specificity and precision for the topS patients (S varying from 1 to 200), considering these definitions:
- True Positive (TP) patients are present in the S most similar patients and present in the list of genetically diagnosed patients provided by the domain expert.
- False Positive (FP) patients are present in the S most similar patients and absent from the list of genetically diagnosed patients
- True Negative (TN) patients are not in the S most similar patients and absent from the list of genetically diagnosed patients
- False Negative (FN) patients are not in the S most similar patients but are present in the list of genetically diagnosed patients
- Recall = TP/(TP + FN); Specificity = TN/(TN + FP); Precision = TP/(TP + FP)

This evaluation is described in Fig. 1.

We repeated this evaluation within the *APDS*, *EBS-DM*, *DEB* and *RETT* cohorts after removing the corresponding diagnosis concepts: "Activated PI3K delta Syndrome, APDS" (CUI C3714976), "Dowling Meara" (CUI C0079295), "dystrophic epidermolysis bullosa" (CUI C0079294) and "Rett syndrome" (CUI C0035372), respectively.

We performed a grid search to determine the hyper parameters: k concepts in common, weight (TFIDF vs. none), concept polarity (all vs. not negated). For this step we combined all the cohorts. We selected the best parameters by using ROC curve analysis.

To visualize the detailed results, we built a ROC curve for each cohort by using the values of the hyper parameters selected during the grid search.

4.2. Second evaluation: removing all the concepts after the date of diagnosis

While Lowe syndrome is usually diagnosed in early infancy (or very first months of life), several rare diseases are diagnosed later in the patient's life, and some others remain undiagnosed for years. For example, APDS is a recently recognized disease, with a mutation of the *PIK3CD* or *PIK3R1* genes identified in 2013. We used *APDS* to evaluate our approach in the situation of late diagnosed diseases. Instead of removing only the "APDS" concept (CUI C3714976) from their list of concepts, we removed all the concepts occurring after the diagnosis. We repeated the same steps as for the first evaluation. Then, we calculated the recall, specificity and precision. The evaluation will be referred as "*APDS-strict*" below.

4.3. Additional evaluations

We calculated the processing time to compute the similarity distance for each Index Patient varying k concepts in common from 0 to 15.

We performed a manual review of the records of the patients returning no true positives. For each cohort, we reviewed manually the records of the 20 patients that were identified as similar the most frequently but were not present in the list of genetically diagnosed patients provided by the expert (False Positives). We also evaluated the association between the number of concepts in the index cases' records and the number of true positive using a linear regression model.

The study was performed in accordance with the precepts of the Declaration of Helsinki and local ethical requirements. The study got approval from the Institutional Review Board #00001072 ref 2016-06-01.

5. Results

In DrWH, the 398,588 patient records are indexed with a total of 9.5 million concepts extracted from clinical narratives, corresponding to 11,037 distinct concepts, (i.e. UMLS Concept Unique Identifier (CUI)). Table 1 describes the cohorts used compared to the population of DrWH. The heterogeneity of the records is illustrated in Tables 2 and 3 with the distribution of these records by hospital departments and by type of reports.

To choose the optimal value of k (the number of concepts in common), the weight (TF-IDF vs. none) and the concept polarity (i.e., negated vs. all) parameters, we built ROC curves calculated on the average recall and specificities of all index patients (Fig. 2) of the 5 cohorts. Fig. 2 shows that, whatever the concepts polarity, the TF-IDF has a huge impact on the ability to display similar patients. The impact of considering all concepts versus exclusion of negated concepts is not as strong as the weight impact, although the recall was slightly better with the use of all concepts. The best

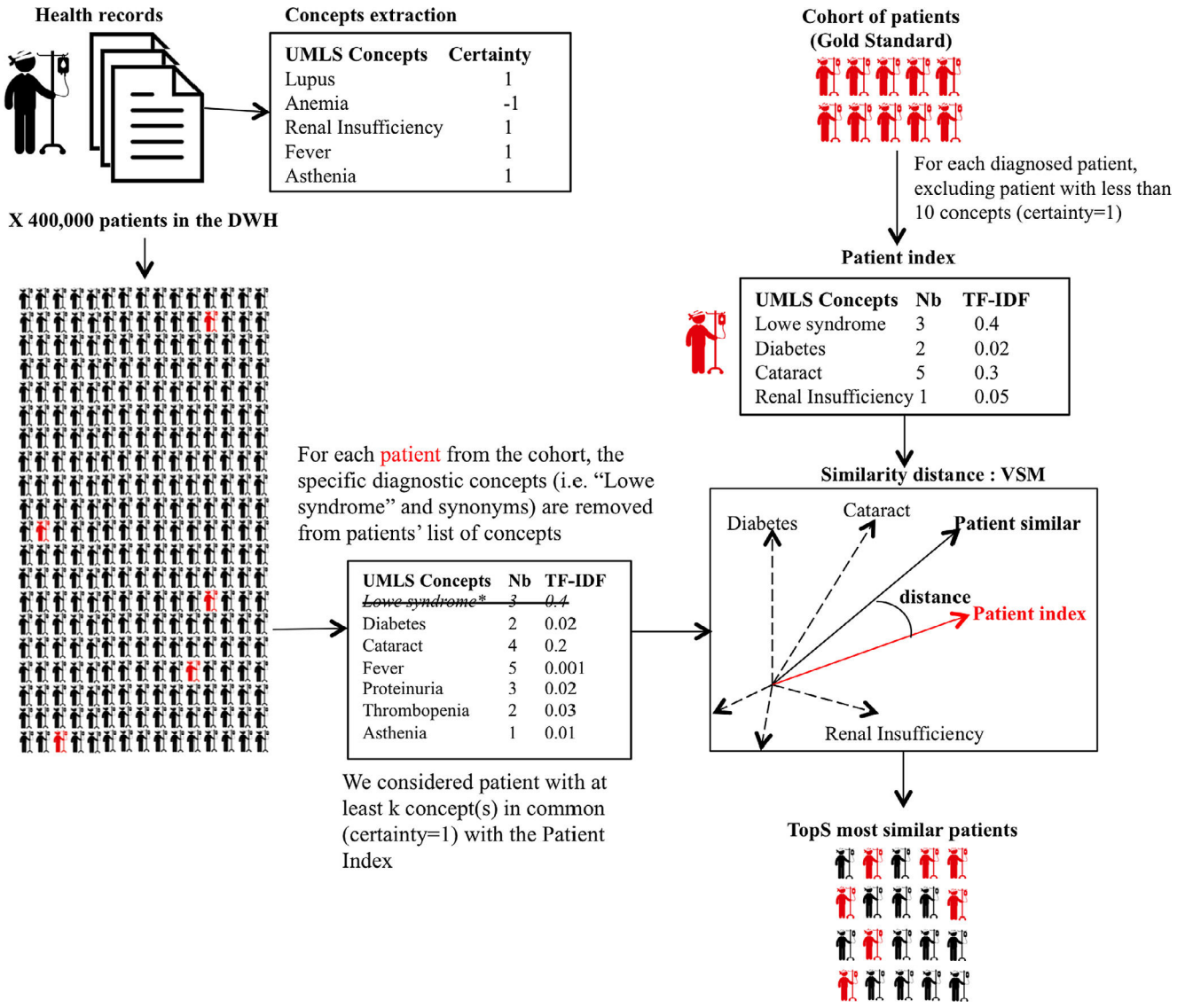


Fig. 1. Evaluation process.

Table 1
Description of the 6 evaluation sets and the population of the data warehouse.

	APDS	*APDS Strict***	*EBS-DM*	*DEB*	*RETT*	*LOWE*	**DWH**
Nb patients	23	23	8	151	87	23	398,588
Nb patients with at least 10 concepts	**21**	**18**	**7**	**103**	**84**	**18**	134,435
Sex ratio (M)	57%	50%	57%	54%	0%	100%	46%
Median Nb reports excluding biological reports*	28 [16-53]	22 [12-44]	15 [5-24]	15 [7-37]	14 [9-24]	43 [18-86]	7 [3-14]
Median Nb concepts*	69 [33-111]	49 [26-76]	63 [23-65]	56 [27-114]	43 [28-67]	58 [39-102]	26 [18-43]
Median Nb concepts excluding negated concepts*	65 [28-89]	59 [35-89]	48 [19-53]	41 [22-91]	36 [24-56]	48 [31-85]	18 [13-30]
Median follow up (years)*	6 [4-8]	4 [3-6]	4 [1-6]	5 [2-7]	6 [5-7]	6 [4-12]	2 [0-5]

*In brackets lower and upper quartile.
** Excluding narrative reports at the time of APDS diagnosis and after.

Table 2
Distribution of the narrative reports by hospital departments for each cohort and in the entire data warehouse (DWH). The numbers in the cells correspond to the percentage of narrative reports in the corresponding cohort. The total number is given in the foot row of the table.

Departments	APDS	APDS strict**	EBD-DM	DEB	RETT	LOWE	DWH
Adult Clinical Hematology	5.8	5.1	0.0	0.0	0.0	0.0	3.7
Adult Nephrology	0.0	0.0	0.0	0.3	0.0	6.5	1.4
Dermatology	2.1	2.7	60.3	56.8	0.1	0.4	1.0
Infectious Diseases	5.1	8.4	0.0	0.0	0.0	0.0	0.4
Maxillofacial Surgery And Stomatology	0.6	0.0	4.1	6.6	2.8	2.3	2.0
Metabolism-Pediatric Neurology	0.9	1.0	7.4	1.8	53.4	2.8	3.7
Pediatric Gastro-Enterology	0.0	0.0	3.3	2.2	9.3	0.8	1.8
Pediatric Immuno-Hematology	71.0	66.1	0.0	0.0	0.0	0.0	1.7
Pediatric Nephrology	0.6	0.8	0.0	0.0	0.1	73.3	3.4
Pediatric Orthopedic Surgery	0.0	0.0	1.7	6.1	7.9	3.1	3.3
Pediatric Radiology	7.0	9.9	1.7	4.3	5.3	2.5	4.4
Pediatric Visceral Surgery	0.1	0.2	5.8	1.3	0.7	1.0	3.1
Physiology	0.0	0.0	4.1	0.2	5.6	0.6	0.7
Pneumo-Allergo Pediatrics	1.2	1.4	2.5	11.2	2.0	0.0	1.1
Other departments	5.6	4.4	9.1	9.2	12.8	6.7	68.3
Total narrative reports	**844**	**513**	**121**	**2375**	**1523**	**961**	**3161978**

** Excluding narrative reports at the time of APDS diagnosis and after.

Table 3
Distribution of the narrative reports by report types for each cohort and in the entire data warehouse (DWH). The numbers in the cells correspond to the percentage of narrative reports in the corresponding cohort. The total number is given in the foot row of the table.

Reports type	APDS	APDS strict**	EBD-DM	DEB	RETT	LOWE	DWH
Medical certificate	0.4	0.2	1.7	1.9	1.9	2.5	0.6
Discharge letter	29.1	32.8	19.8	24.8	17.3	24.8	17.5
Pathology report	0.5	0.0	0.0	1.4	0.3	0.1	0.6
Consultation	28.8	28.3	39.7	23.2	52.5	34.7	20.4
EEG	0.0	0.0	4.1	0.1	5.2	0.3	0.5
Ambulatory care	23.9	14.8	13.2	18.0	1.5	3.5	2.2
Hospitalization	6.9	9.7	16.5	22.1	13.2	5.2	5.5
Radiology	7.6	10.1	2.5	5.3	4.6	6.0	9.3
Operative	2.8	4.1	2.5	2.8	2.1	0.5	2.7
Staff	0.0	0.0	0.0	0.1	0.3	0.1	0.1
Emergency	0.0	0.0	0.0	0.3	1.1	0.0	1.0
Prescription	0.0	0.0	0.0	0.0	0.0	22.3	0.6
Lab test reports not included in the evaluation	0.0	0.0	0.0	0.0	0.0	0.0	39.0
Total	**844**	**513**	**121**	**2375**	**1523**	**961**	**3161978**

** Excluding narrative reports at the time of APDS diagnosis and after.

value for the parameter "k concepts in common" was 8 with TF-IDF and all the concepts (at top30 similar patients, average recall = 0.20 ± 0.11, precision = 0.51 ± 0.32).

For all further analyses, we kept k equal to 8, TF-IDF weighted and including negated concepts. Fig. 3 shows 6 ROC curves for the five evaluation sets and one all together, As suggested by the six physician experts (NBB, SK, HRS, SR, SF, AF, AM), we considered the top 30 similar patients as they would not consider results beyond this limit. In the 30 first most similar patients, a diagnosed patient returned an average of 51% of true positive patients (*DEB* 0.75 ± 0.24, *LOWE* 0.18 ± 0.9, *RETT* 0.44 ± 0.24, *EBS-DM* 0.08 ± 0.04, *APDS* 0.09 ± 0.09). The precision for *EBS-DM* and *APDS* were low because of the low number of patients in these cohorts (7 and 21 respectively).

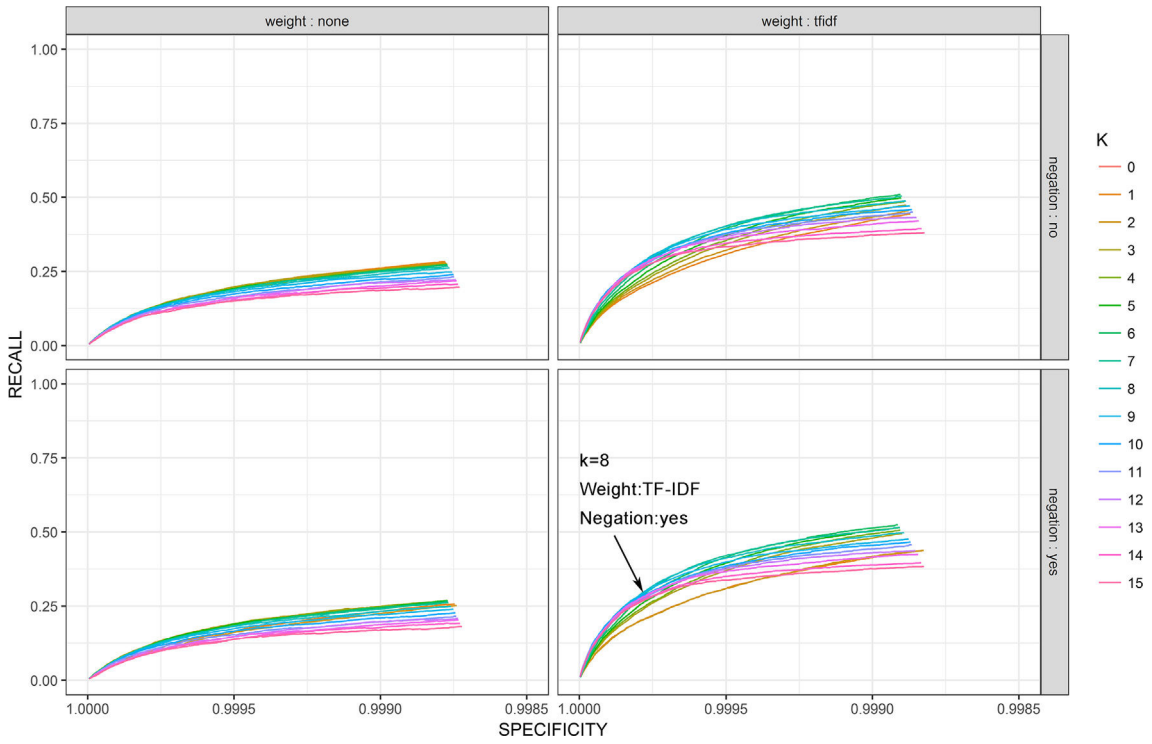


Fig. 2. ROC curve calculated on the average recall and specificity of all index patients, varying k (0–15), with weight = TFIDF or without weight, and including negated concepts or not. We varied n most similar patients from 1 to 200.

The recall, precision and specificity for the *APDS-strict* were close from those from *APDS* (using k = 8, including negated concepts and TF-IDF weighted). The average recall at the Top30 was 0.16 (±0.11) and the precision was 0.10 (±0.07).

The percentages of index patients returning at least one true positive in the Top30 similar patients were 94% for *LOWE*, 97% for *DEB*, 86% for *APDS*, 71% for *EBS-DM* and 99% for *RETT*. After manual review of the records of the index patients returning no true positive similar patients, we identified two major causes: (i) number of concepts <20 in the index patient records, (ii) differential diagnosis of the index patient (e.g., suggesting CD40 ligand deficiency rather than APDS syndrome). We identified a statistically significant association between the number of concepts in the index patient's records and the number of true positive similar patients (p-value = 1.03e−08). For example, in the *LOWE* cohort, two patients had very few concepts associated with their Lowe syndrome because they came at Necker-*Enfants Malades* Hospital for orthopedic surgery and had no follow-up with a specialist of this rare disease.

The manual review of patients identified as similar using our method but not included in the list of genetically diagnosed patients (False Positives) showed two major causes: patients had phenotypes that were compatible with the index patient phenotype but explained by another cause (e.g. CD40 ligand deficiency instead of APDS). CD40 ligand deficiency and APDS are both primary immunodeficiencies associated with hyper-IgM syndrome [41]. In addition, for some false positive patients, physicians had requested genetic testing of known causative mutations, but the results of testing came back negative. Finally, some patients did not have a specific diagnosis (e.g. "not otherwise specified immunodeficiency") or the results regarding the suspected mutations were not available yet.

Regarding the processing time, as seen in Fig. 4, it depends on the k parameter (number of concepts in common). Above 8 concepts in common, the processing time was stable, around 12 s using Intel® Xeon® E5603 processor (1,60 GHz – 4 Mo dual core 80 W) and DDR3 16 GB random access memory. We considered this time reasonable for a user to get a result in an interactive web interface.

6. Discussion

Our method allowed the identification of similar patients based on the concepts extracted from the free-text reports. Additionally, this method demonstrated its ability to identify undiagnosed patients with similar phenotypic features. Moreover, in the context of rare diseases, with prevalence ranging from 1/1.7 Millions to 1/15,000, and using hospital cohorts, we demonstrated the ability of our algorithm to identify clinically similar cases in a clinical data warehouse. Despite a rather low precision, Necker geneticists considered the system helpful. When a new mutation is discovered on a patient, geneticists consider that similar patients should be looked for in order to explore their genotypes. Finding even only one more patient is already useful and significant progress for research and patient care.

The "k concepts in common" hyper parameter has been tuned by using the same cohorts as those used for the evaluation. Indeed, we were constrained by the number of validated cohorts available at Necker Hospital. Therefore we could not divide the data into training and validation sets to evaluate our method. Consequently the performance is likely overstated.

Our method can be adapted for use in any dataset, in any language, without any training. The addition of TF-IDF scores to weight the different concepts increased the recall. TF-IDF scores are easily implemented and very scalable. In our setting, the method was based on concepts extracted from full-text reports, but it could also be extended to ICD codes or other structured data such as biological test results. We integrated lexical methods in the ETL pipeline to detect negation, family contexts and the semantic

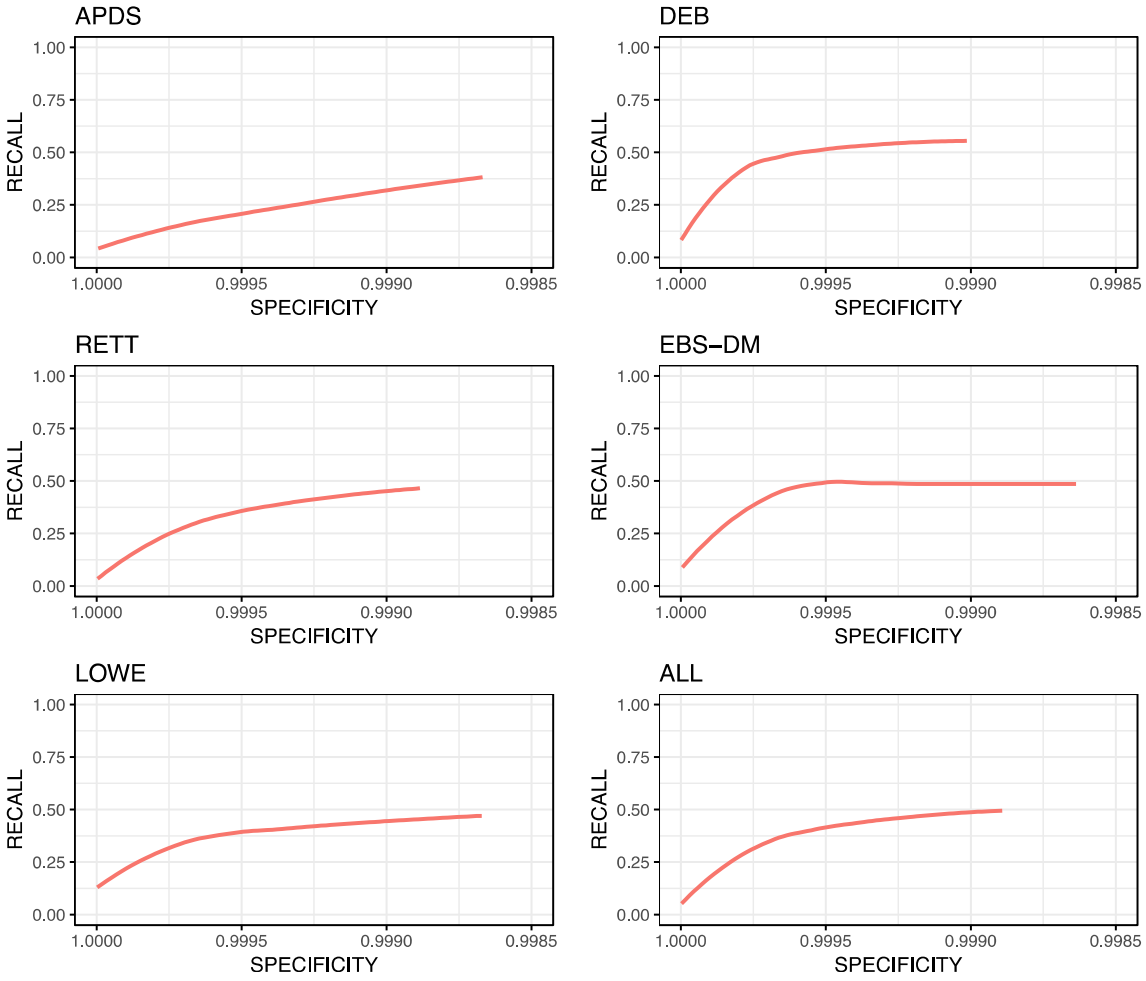


Fig. 3. For each cohort, ROC Curve, average recall and specificity of all index patients, k = 8, all concepts and weight = tf-idf – 1–200 similar patients.

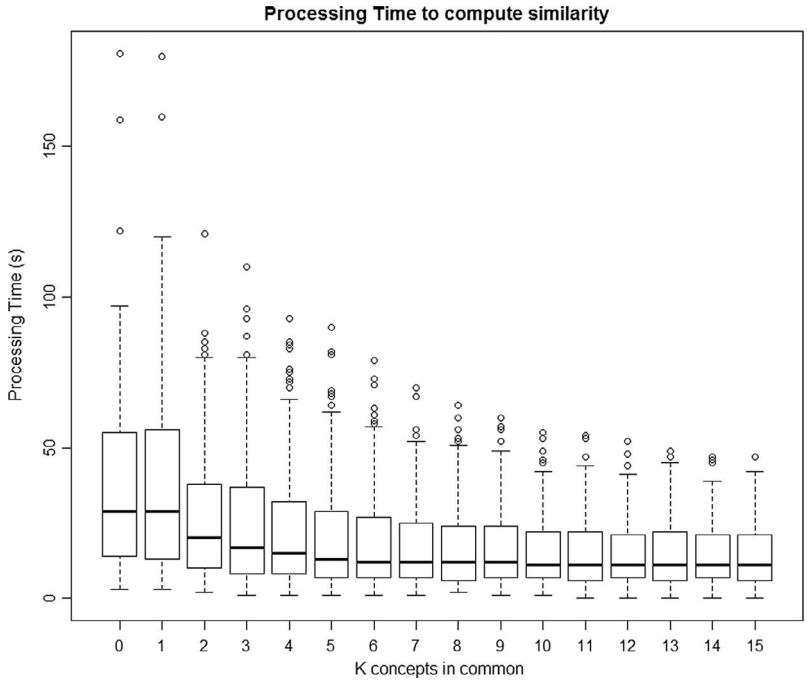


Fig. 4. Median Processing Time (s) to compute similarity on the 233 patients, varying k concepts in common from 0 to 15. For example, k = 8, median processing time = 12 s.

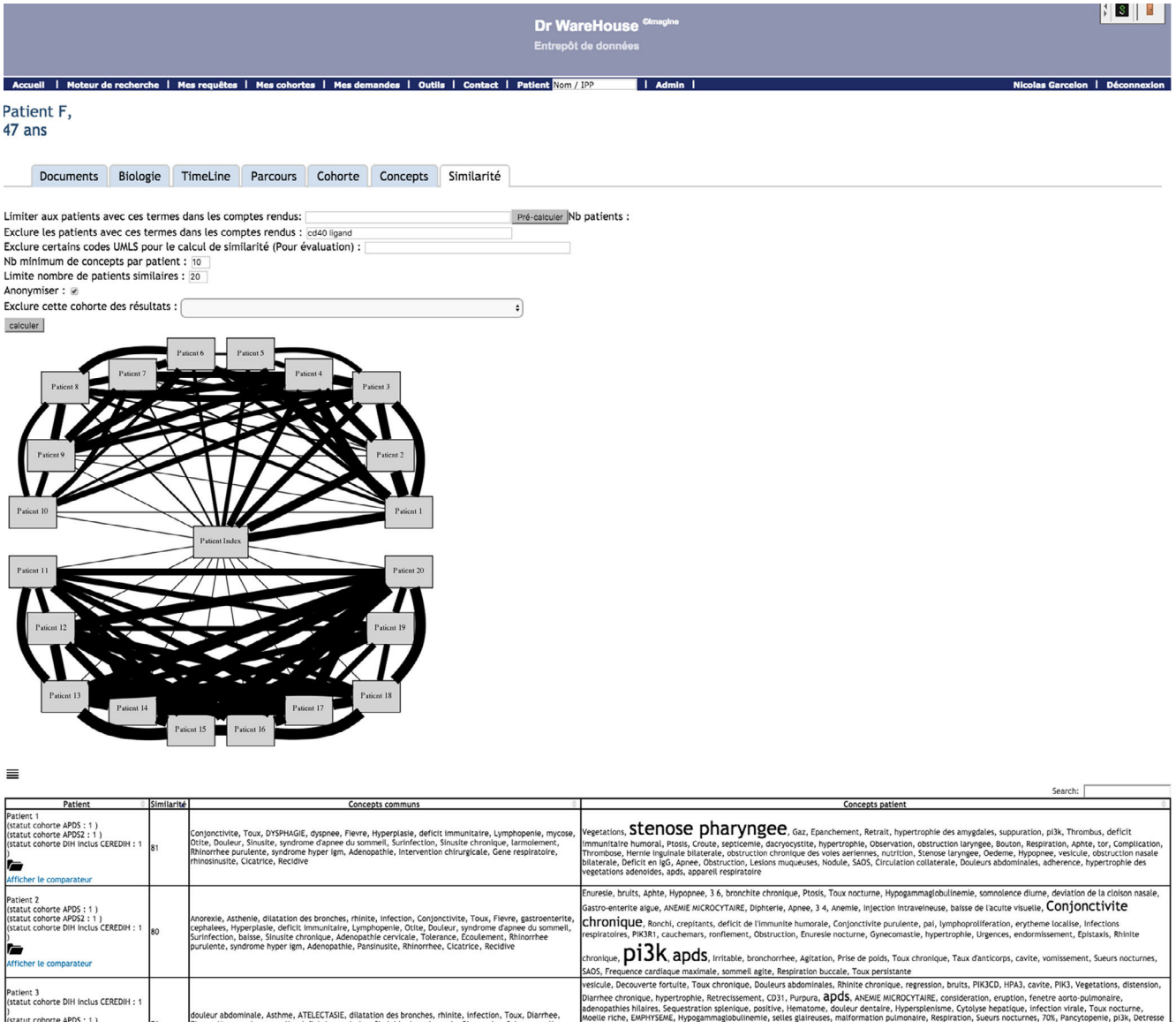


Fig. 5. visualization of a patient in Dr. Warehouse. Similarity tab: it displays the 20 most similar patients. The network in the middle displays the links between each similar patient. The width of links depends of the similarity. The table lists each patient and displays for each similar patient the concepts in common with the Index Patient and concepts only in similar patient.

resources of the UMLS to enhance the precision of the concept extraction [29].

The vector space model is a recognized, robust and easy to implement model based on linear algebra. As such, this model provides explicit and interpretable results for the clinicians. Compared to Boolean queries, VSM enables partial matching based on the concepts of the index patient. In addition, the processing time of this method – which is currently performed within seconds – is compatible with real time usage in a web interface.

From a clinical point of view, finding relevant cases may be useful and enable the recruitment of patients with no diagnosis yet for genetic testing, screening tests or clinical studies. Boolean queries require a finite list of criteria validated by disease experts. They may not be sensitive enough and could omit atypical cases. For example, a Boolean query on DrWH using the terms "proximal tubulopathy" and "hypotonia" and "cataract" to identify Lowe syndrome patients [34] returned only 3 patients. In comparison, our methods identified on average 6 patients. On the other hand, some syndromes such as APDS and Rett syndrome are defined by unspecific criteria, resulting in a broad spectrum of results when using Boolean queries. Our approach allows more flexibility by using continuous similarity scores instead of binary decisions.

This work has some limitations that are inherent to the use of a VSM. First, the VSM assumes that terms are statistically independent. In the medical context, this hypothesis is rarely supported. Therefore, this limitation may interfere with the quality of the results. Another limitation lies in the vector representation of concepts. In this representation, each concept, as a support vector, supports a dimension of the space, and all the dimensions are at equal distance from each other. If two concepts are semantically close, this proximity is not taken into account in this type of representation. Techniques based on distributional semantic models, namely word embedding, as well as semantic methods based on terminologies and ontologies address this issue [42]. These models take into account the associations and the semantic relations between concepts to compute a distance between concepts and enhance the estimation of similarity between patients.

The quality and the completeness of the data in the data warehouse also influenced the quality of these results. Better model performance would possibly be observed with additional data

sources, e.g., ICD codes, and laboratory test results. Moreover, the heterogeneity of the phenotypes associated with the same mutations as well as the phenotypic proximity to other diseases' phenotypes also played a role in the differences in accuracy observed between the cohorts. This is the case with CD40L and APDS, Lowe and Zellweger syndrome [43], Lowe and Smith-Lemli-Opitz syndrome, EBD-DM and Biotin deficiency [44]).

We integrated this similarity tool in Dr. Warehouse® (Fig. 5). Clinicians may use this tool to help identifying potential diagnoses and genetic mutations. Additionally, researchers and physicians may use this tool to identify patients in the data warehouse who could benefit from genetic testing, especially when new genetic causes are discovered. Similarity metrics applied to rare disease repositories offer new perspectives in a translational context that may help to recruit patients for research and reduce the length of the diagnostic journey.

7. Conclusion

Time to diagnosis for rare disease patients and undiagnosed diseases are crucial health issues that require increased efforts in the scientific and healthcare fields [45]. The tool described here can help to identify new patients with rare diseases. We consider this method in a data-driven approach to diagnose undiagnosed patient by using case-based reasoning on similar diagnosed patients [46].

Ethics approval

We got an ethical approval by the French IRB CPP Il-de-France II (IRB registration number 00001072) registered under reference 2016-06-01.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

None.

References

[1] The Shire Rare Disease Impact Report (2013 – US and UK Population). <https://globalgenes.org/wp-content/uploads/2013/04/ShireReport-1.pdf> (accessed 26 Sep 2016).
[2] 5 Questions: Euan Ashley on Diagnosing the Undiagnosable. <http://med.stanford.edu/news/all-news/2015/09/5-questions-euan-ashley-on-diagnosing-the-undiagnosable.html> (accessed 26 Sep 2016).
[3] I. Danciu, J.D. Cowan, M. Basford, et al., Secondary use of clinical data: the Vanderbilt approach, J. Biomed. Inform. 52 (2014) 28–35, http://dx.doi.org/10.1016/j.jbi.2014.02.003.
[4] S.N. Murphy, G. Weber, M. Mendis, et al., Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J. Am. Med. Inform. Assoc. JAMIA 17 (2010) 124–130, http://dx.doi.org/10.1136/jamia.2009.000893.
[5] H.J. Lowe, T.A. Ferris, P.M. Hernandez, et al., STRIDE–an integrated standards-based translational research informatics platform, AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp. 2009 (2009) 391–395.
[6] J. Sun, F. Wang, J. Hu, et al., Supervised patient similarity measure of heterogeneous patient records, SIGKDD Explor. Newsl. 14 (2012) 16–24, http://dx.doi.org/10.1145/2408736.2408740.
[7] G. Salton, M. Lesk, Computer evaluation of indexing and text processing, J. ACM 15 (1968) 8–36.
[8] G. Salton, The SMART Retrieval System—Experiments in Automatic Document Processing, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.
[9] G. Salton, Automatic Information Organization and Retrieval, McGraw Hill Text, 1968.
[10] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1972) 11–21.
[11] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.
[12] J. Zobel, A. Moffat, R. Wilkinson, et al., Efficient retrieval of partial documents, Inf. Process. Manag. 31 (1995) 361–377, http://dx.doi.org/10.1016/0306-4573(94)00052-5.
[13] D.L. Lee, H. Chuang, K. Seamons, Document ranking and the vector-space model, IEEE Softw. 14 (1997) 67–75, http://dx.doi.org/10.1109/52.582976.
[14] W.M. Shaw, R. Burgin, P. Howell, Performance standards and evaluations in IR test collections: vector-space and other retrieval models, Inf. Process. Manag. 33 (1997) 15–36, http://dx.doi.org/10.1016/S0306-4573(96)00044-1.
[15] C.J. Crouch, A. Mahajan, A. Bellamkonda, Flexible retrieval based on the vector space model, in: N. Fuhr, M. Lalmas, S. Malik (Eds.), Advances in Xml Information Retrieval, Springer-Verlag, Berlin, 2005, pp. 292–302.
[16] J. Zobel, A. Moffat, Inverted files for text search engines, ACM Comput. Surv. 38 (2006), http://dx.doi.org/10.1145/1132956.1132959.
[17] M. Kwak, G. Leroy, J.D. Martinez, et al., Development and evaluation of a biomedical search engine using a predicate-based vector space model, J. Biomed. Inform. 46 (2013) 929–939, http://dx.doi.org/10.1016/j.jbi.2013.07.006.
[18] K. Golub, Automated subject classification of textual web documents, J. Doc. 62 (2006) 350–371, http://dx.doi.org/10.1108/00220410610666501.
[19] L. Xie, G. Li, M. Xiao, et al., Novel classification method for remote sensing images based on information entropy discretization algorithm and vector space model, Comput. Geosci. 89 (2016) 252–259, http://dx.doi.org/10.1016/j.cageo.2015.12.015.
[20] L. Jing, M.K. Ng, J.Z. Huang, Knowledge-based vector space model for text clustering, Knowl. Inf. Syst. 25 (2010) 35–55, http://dx.doi.org/10.1007/s10115-009-0256-5.
[21] P.-C. Lee, H.-N. Su, T.-Y. Chan, Assessment of ontology-based knowledge network formation by Vector-Space Model, Scientometrics 85 (2010) 689–703, http://dx.doi.org/10.1007/s11192-010-0267-8.
[22] I. Santos, C. Laorden, B. Sanz, et al., Enhanced topic-based vector space model for semantics-aware spam filtering, Expert Syst. Appl. 39 (2012) 437–444, http://dx.doi.org/10.1016/j.eswa.2011.07.034.
[23] P. Castells, M. Fernandez, D. Vallet, An adaptation of the vector-space model for ontology-based information retrieval, IEEE Trans. Knowl. Data Eng. 19 (2007) 261–272, http://dx.doi.org/10.1109/TKDE.2007.22.
[24] I.N. Sarkar, A vector space model approach to identify genetically related diseases, J. Am. Med. Inform. Assoc. JAMIA 19 (2012) 249–254, http://dx.doi.org/10.1136/amiajnl-2011-000480.
[25] J. Lee, D.M. Maslove, J.A. Dubin, Personalized mortality prediction driven by electronic medical data and a patient similarity metric, PloS One 10 (2015) e0127428, http://dx.doi.org/10.1371/journal.pone.0127428.
[26] B. Campillo-Gimenez, N. Garcelon, P. Jarno, et al., Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France, Stud. Health Technol. Inform. 192 (2013) 572–575.
[27] S.T. Rosenbloom, J.C. Denny, H. Xu, et al., Data from clinical notes: a perspective on the tension between structure and flexible documentation, J. Am. Med. Inform. Assoc. JAMIA 18 (2011) 181–186, http://dx.doi.org/10.1136/jamia.2010.007237.
[28] P. Raghavan, J.L. Chen, E. Fosler-Lussier, et al., How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?, AMIA Jt Summits Transl. Sci. Proc. AMIA Summit. Transl. Sci 2014 (2014) 218–223.
[29] N. Garcelon, A. Neuraz, V. Benoit, 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, J. Am. Med. Inform. Assoc. JAMIA. http://dx.doi.org/10.1093/jamia/ocw144. Published Online First: 20 October 2016.
[30] W.W. Chapman, W. Bridewell, P. Hanbury, et al., Evaluation of negation phrases in narrative clinical reports, Proc. AMIA Symp. 2001 (2001) 105–109.
[31] D. Chu, J.N. Dowling, W.W. Chapman, Evaluating the effectiveness of four contextual features in classifying annotated clinical conditions in emergency department reports, AMIA Annu. Symp. Proc. AMIA Symp. (2006) 141–145.
[32] A.E. Guttmacher, F.S. Collins, R.H. Carmona, The family history–more important than ever, N. Engl. J. Med. 351 (2004) 2333–2336, http://dx.doi.org/10.1056/NEJMsb042979.
[33] Orphanet. <http://www.orpha.net/consor/www/cgi-bin/index.php?lng=FR> (accessed 22 Feb 2017).
[34] M. Loi, Lowe syndrome, Orphanet. J. Rare Dis. 1 (2006) 16, http://dx.doi.org/10.1186/1750-1172-1-16.
[35] A. Hovnanian, L. Hilal, C. Blanchet-Bardon, et al., Recurrent nonsense mutations within the type VII collagen gene in patients with severe recessive dystrophic epidermolysis bullosa, Am. J. Hum. Genet. 55 (1994) 289–296.
[36] P.H.L. Schuilenga-Hut, P.v.d. Vlies, M.F. Jonkman, et al., Mutation analysis of the entire keratin 5 and 14 genes in patients with epidermolysis bullosa simplex and identification of novel mutations, Hum. Mutat. 21 (2003) 447, http://dx.doi.org/10.1002/humu.9124.
[37] R.E. Amir, I.B. Van den Veyver, M. Wan, et al., Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2, Nat. Genet. 23 (1999) 185–188, http://dx.doi.org/10.1038/13810.
[38] I. Angulo, O. Vadas, F. Garçon, et al., Phosphoinositide 3-kinase δ gene mutation predisposes to respiratory infection and airway damage, Science 342 (2013) 866–871, http://dx.doi.org/10.1126/science.1243292.
[39] M.-C. Deau, L. Heurtier, P. Frange, et al., A human immunodeficiency caused by mutations in the PIK3R1 gene, J. Clin. Invest. 124 (2014) 3923–3928, http://dx.doi.org/10.1172/JCI75746.
[40] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., 1986. <http://lyle.smu.edu/~mhd/8337sp07/salton.pdf> (accessed 17 Sep 2015).

[41] T.I. Coulter, A. Chandra, C.M. Bacon, et al., Clinical spectrum and features of activated phosphoinositide 3-kinase δ syndrome: a large patient cohort study, J. Allergy Clin. Immunol. http://dx.doi.org/10.1016/j.jaci.2016.06.021. Published Online First: 16 July 2016.
[42] T. Mikolov, K. Chen, G. Corrado, et al., Efficient Estimation of Word Representations in Vector Space. ArXiv13013781 Cs. <http://arxiv.org/abs/1301.3781> (accessed 24 Sep 2016). Published Online First: 16 January 2013.
[43] V. Castrodale, The hypotonic infant: case study of central core disease, Neonatal Netw. NN 22 (2003) 53–59, http://dx.doi.org/10.1891/0730-0832.22.1.53.
[44] S. Jackson, L.T. Nesbitt, Differential Diagnosis for the Dermatologist, Springer Science & Business Media, 2012.
[45] D. Taruscio, S.C. Groft, H. Cederroth, et al., Undiagnosed Diseases Network International (UDNI): white paper for global actions to meet patient needs, Mol. Genet. Metab. 116 (2015) 223–225, http://dx.doi.org/10.1016/j.ymgme.2015.11.003.
[46] J. Frankovich, C.A. Longhurst, S.M. Sutherland, Evidence-based medicine in the EMR era, N. Engl. J. Med. 365 (2011) 1758–1759, http://dx.doi.org/10.1056/NEJMp1108726.

5. Résultats

5.1. L'entrepôt de données de l'hôpital Necker Enfants Malades

Au 31 aout 2017, l'entrepôt de données de l'hôpital Necker Enfants Malades contient 489 705 patients et 3 923 597 documents. Nous avons intégré 21 sources de données différentes décrites dans le Tableau 2, dont 11 sont des sources de données institutionnelles (en gras dans le tableau), et 10 sont des sources de données locales ou spécifiques à un service. Ces bases de données n'ont pas systématiquement l'identifiant patient hospitalier.

Tableau 2 : description des sources de données dans l'entrepôt de données de l'hôpital Necker. Les sources suivies d'une étoile sont des sources institutionnelles avec les identifiants hospitaliers des patients.

APIX*	**Logiciel d'anatomopathologie**
ASTRAIA	Logiciel des échographies anténatales
Base IRM	Base de données de recherche IRM cérébraux et génétique
Base NPH	Base de données de recherche sur les néphronophtises
COMMUN_CARDIO	Dossier contenant les CR de cardiologie pédiatrique (avant le DPI)
COMMUN_CHIRVISC	Dossier contenant les CR de chirurgie viscérale (avant SUSIE)
COMMUN_NEPHROPED	Dossier contenant les CR de néphrologie pédiatrique (avant SUSIE)
COMMUN_ORLTROUSSEAU	Dossier contenant les CR d'ORL de Trousseau avant fusion avec Necker
COMMUN_TRAUMA_CMI	Dossier contenant les CR de traumatologie
COMMUN_TRAUMA_CRH	Dossier contenant les CR de traumatologie
DIAMIC*	**Logiciel de gestion d'anatomopathologie**
DIAMMG*	**Logiciel de gestion des données génétiques**
Foetopath	Dossier contenant les CR de foetopathologie
KDOS*	**Logiciel de gestion des CR des réunions de concertation pluridisciplinaire**
MEDIWEB*	**Dossier contenant les CR médicaux**
ORBIS*	**Dossier patient informatisé**
PACS*	**Logiciel des données d'imagerie**
PMSI_DIAG*	**Diagnostic CIM10 du PMSI**
RADOS*	**Logiciel de création des CR d'imagerie avant le PACS**
STARE*	**Logiciel de gestion des résultats biologiques**
SUSIE*	**Logiciel de gestion des CR médicaux avant le DPI**

La Figure 2 illustre la disparité des données dans le temps pour chacune des sources de 1996 à 2017. Sur les 21 sources, seules 5 d'entre elles fournissent des données structurées (
Tableau 3). Les 16 autres sources de données ne fournissent que des comptes rendus en texte libre.

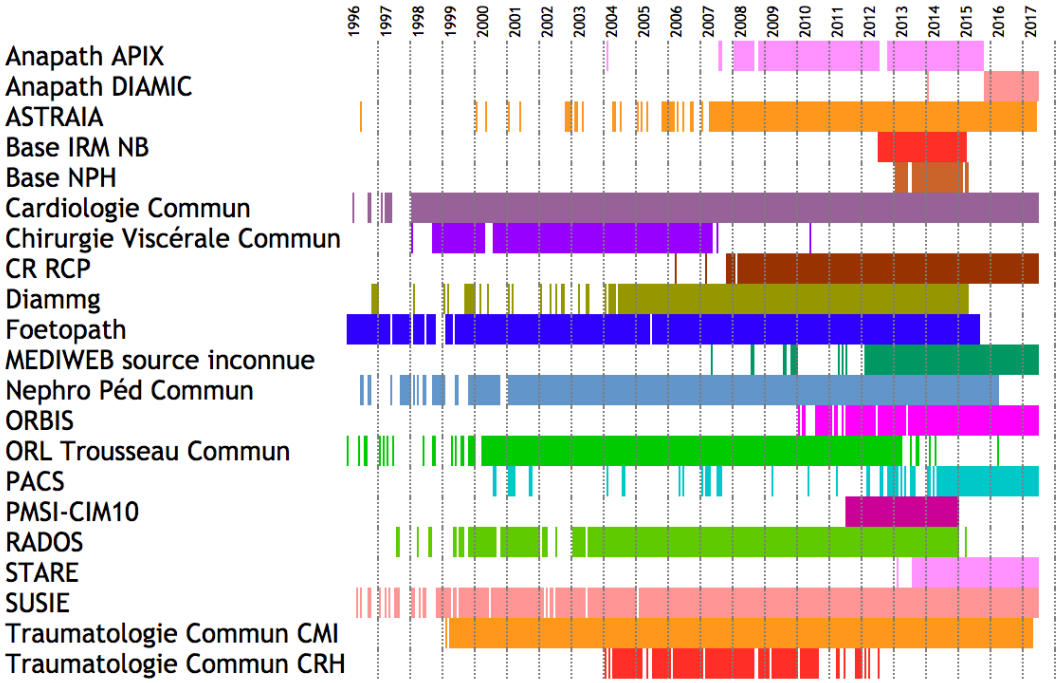


Figure 2 : Distribution au cours du temps des documents par source de données

Tableau 3 : Distribution du nombre de documents et patients par source de données intégrée dans l'entrepôt de données de l'hôpital Necker

Sources	Nb patients	Nb Documents	Nb Concepts extraits	Nb Données codées	Périodes
APIX	10 094	12 955	61 995	-	2004-2015
ASTRAIA	53 324	345 909	1 501 796	13 239 529	1996-2017
Base IRM	4 134	4 210	6 786	-	2012-2015
Base NPH	417	417	3 000	197	2013-2015
COMMUN_CARDIO	55 515	217 668	1 163 619	-	1996-2017
COMMUN_CHIRVISC	1 967	4 271	23 263	-	1996-2010
COMMUN_NEPHROPED	7 376	70 714	287 963	-	1996-2016
COMMUN_ORLTROUSSEAU	17 246	22 642	97 454	-	1996-2016
COMMUN_TRAUMA_CMI	2 433	2 546	7 096	-	1999-2017
COMMUN_TRAUMA_CRH	227	232	2 693	-	2004-2012
DIAMIC	9 380	12 918	67 862	-	2014-2017
DIAMMG	3 604	5 363	19 823	5 363	1996-2015
Foetopath	5 974	6 849	71 066	-	1996-2015
KDOS	2 662	4 908	60 830	-	2006-2017
MEDIWEB	20 381	33 331	62 446	-	2007-2017
ORBIS	103 763	374 737	3 697 273	-	2010-2017
PACS	46 698	90 981	465 401	-	2000-2017
PMSI_DIAG	84 922	197 404	456 844	679 435	2011-2014
RADOS	89 469	229 253	1 114 613	-	1997-2015
STARE	121 355	1 415 632	-	21 412 363	2013-2017
SUSIE	227 464	870 657	9 526 115	-	1996-2017
Total	**489 705**	**3 923 597**	**18 697 938**	**35 336 887**	**1996-2017**

Un total de 18 697 938 concepts a été extrait des comptes rendus cliniques. 96% ont été associés au patient et 4% aux antécédents familiaux. Parmi les 96%, 72% sont considérés comme non négatifs (Tableau 4).

Tableau 4 : Nombre de concepts extraits par contexte et niveau de certitude

Contexte / Certitude	Négation	Non négation	Nb concepts Total
Antécédents familiaux	179 938	522 009	701 947
Patient	5 007 517	12 988 474	17 995 991
Nb concepts Total	**5 187 455**	**13 510 483**	**18 697 938**

En moyenne il y a 8 documents par patient avec un patient qui a 1 665 comptes rendu. Cela fait une moyenne de 42 concepts par patient et une médiane de 5 concepts (Tableau 5).

Tableau 5 : Moyenne (écart type), médiane (interquartile bas et haut), minimum et maximum du nombre de documents par patient, du nombre de concepts par patient, du nombre de données codées par patient, et du nombre de concepts par document

	Moyenne	Médiane	Minimum	Maximum
Nb Documents par patient	8,01 (+/- 25,78)	3 (1-6)	1	1 665
Nb Concepts par patient	41,88 (+/- 119,36)	13 (5-35)	1	6 293
Nb Données codées par patient	167,11 (+/- 469,49)	41 (7-154)	1	21 790
Nb Concepts par document	8,29 (+/- 10,41)	5 (3-10)	1	317

Nous avons pu établir 37 111 liens de parenté dans l'entrepôt de données (36 805 relations enfant-mère, 289 relations enfant-père).

5.2. Dr Warehouse – le logiciel

L'ensemble des méthodes et algorithmes décrits dans ce manuscrit a été intégré dans un logiciel diffusé sous licence open source. Cette diffusion a été associée à la soumission d'un article au Journal of Biomedical Informatics. Suite à son déploiement partiel en janvier 2017 sur l'hôpital Necker Enfants Malades, nous avons réalisé une enquête de satisfaction en août 2017, basée sur l'enquête réalisée par l'université du Michigan sur l'utilisation d'Emerse (Hanauer et al., 2015). Nous décrivons les résultats de cette enquête ainsi que les statistiques d'utilisation du logiciel pendant les 6 premiers mois (janvier à juin 2017).

Le logiciel est en téléchargement libre sur une plateforme GIT : https://github.com/imagine-bdd/DRWH

Le logiciel est accessible en démonstration sur internet avec des données générées à partir des résumés PubMed : https://imagine-plateforme-bdd.fr/dwh_pubmed/

A clinician friendly data warehouse oriented toward narrative reports: Dr Warehouse

Nicolas Garcelon, MSc[1,2], Antoine Neuraz, MD[2,3], Rémi Salomon, MD PhD [1,4],Hassan Faour[1], Vincent Benoit, PhD[1], Arthur Delapalme[1], Arnold Munnich, MD PhD[1,5], Anita Burgun, MD PhD[2,3], Bastien Rance, PhD[2,6]

[1]*Institut Imagine, Paris Descartes Université Paris Descartes-Sorbonne Paris Cité, Paris, France;*
[2]*INSERM, Centre de Recherche des Cordeliers, UMR 1138 Equipe 22, Université Paris Descartes, Sorbonne Paris Cité, Paris, France;*
[3]*Department of Medical informatics, Hôpital Necker-Enfant Malades, Assistance Publique des Hôpitaux de Paris, Paris, France*
[4]*Service de Néphrologie Pédiatrique, Hôpital Necker-Enfants Malades, Assistance Publique -Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France*
[5]*Département de génétique médicale, Hôpital Necker-Enfants Malades, Assistance Publique -Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France. Centre de Référence des Maladies Osseuses Constitutionnelles, INSERM UMR 1163, Laboratoire de bases moléculaires et physiopathologiques de l'ostéochondrodysplasie, Paris Descartes-Sorbonne Paris Cité University, AP-HP, Institut Imagine, 75015 Paris, France*
[6]*Hôpital Européen Georges Pompidou, Assistance Publique -Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, Franc*

Competing interests: None.
Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Ethics approval: We got an ethical approval by the French IRB CPP Il-de-France II (IRB registration number 00001072) registered under reference 2016-06-01.

Keywords
Software, Computational biology, Method, Data warehouse, Rare diseases, Electronic Health Records, Information storage and retrieval, Text-mining

Abstract

Introduction.
Clinical data warehouses are often oriented toward integration and exploration of coded data. However narrative reports are of crucial importance for translational research. This paper describes Dr Warehouse®, an open source data warehouse oriented toward clinical narrative reports and designed to support clinicians' day-to-day use.
Method.
Dr Warehouse relies on an original database model to focus on documents in addition to facts. Besides classical querying functionalities, the system provides an advanced search engine and Graphical User Interfaces adapted to the exploration of text. Dr Warehouse is dedicated to translational research with cohort recruitment capabilities, high throughput phenotyping and patient centric views (including similarity metrics among patients). These features leverage Natural Language Processing based on the extraction of UMLS® concepts, as well as negation and family history detection.
Results.
A survey conducted after 6 months of use at the Necker Children's Hospital shows a high rate of satisfaction among the users (96.6%). During this period, 122 users performed 2,837 queries, accessed 4,267 patients' records and included 36,632 patients in 131 cohorts.
The source code is available at this github link https://github.com/imagine-bdd/DRWH.
A demonstration based on PubMed abstracts is available at https://imagine-plateforme-bdd.fr/dwh_pubmed/.

1. Background and significance

Built on top of EHRs, Clinical Data Warehouses (CDWs) enable collection and secondary use of healthcare data for many purposes, including research. CDWs have been used for many types of applications covering all the realms of medicine: Phenome-wide analysis,[1–4] record mining,[5–10] epidemiological surveillance, pharmacovigilance, etc.

CDWs are often used to leverage structured data (e.g. billing codes, procedures, laboratory results, treatments). However, despite important effort toward the standardization of data collection, a large part of the medical knowledge remains embedded in free-text clinical narratives (including discharge summaries, letters…). Indeed, *Raghavan et al.* showed that solving the inclusion criteria for clinical studies required the free-text reports in 60 to 80% of the cases.[11] Furthermore, the use of free text enables clinicians to report more subtle information such as doubts, the absences of indicators, or diagnostic hypotheses [8,12]. This is especially important in the context of rare diseases for which free text remains the ideal means of preserving the phenotypical richness of a patient case, and focuses the patient-doctor relationship around narrative medicine [13]. In addition, terminologies are susceptible to fast and constant evolution [14] making it difficult to maintain terminological alignments and annotations. Medical IT teams or data scientists must therefore develop methods that make possible to reuse the data produced within the care process for the purposes of research, teaching and management [15].

Related Work

There is a large body of literature around clinical data warehousing. However, only a limited number of systems have been both published and distributed as open-source system. i2b2 [16,17] the framework from the eponym NIH-funded National Center for Biomedical Computing based at Partners HealthCare System in the US, is a widely used system worldwide. More recently, the OHDSI [18] initiative (formerly the Observational Medical Outcomes Partnership - OMOP) has contributed to the standardization of warehouses by building and maintaining the CDM (Common Data Model) [19], adopted by many partners over the world. Both systems adopted an Entity Attribute Value model (EAV). I2b2 uses a star schema with facts stored in a single table. The CDM model split the table of facts into domain-oriented tables, but the philosophy is overall similar. Recently, i2b2 released an OHDSI-compatible version. If it is possible to store free text in both of the systems, they were not designed for this usage. Their interfaces are mainly dedicated to the interrogation of coded information.

Numerous CDW solutions have been proposed, including STRIDE [20], the Vanderbilt solution [21], The Enterprise Data Trust at Mayo Clinic [22], radBank developed by the Stanford Medical Informatics Department [23], DW4TR developed by the Windber Research Institute [24], METEOR [25], Starmaker [7], Consore [26], SMEYEDAT (University Eye Hospital of Munich, Germany) [27]. These CDW are not distributed, and their ability to handle and query clinical narratives remains unclear.

EMERSE (the Electronic Medical Record Search Engine) [28], developed by the University of Michigan, is designed to enable the interrogation of documents within an EHR. EMERSE was designed to be used by users with no expertise in IT or informatics. EMERSE is not, strictly speaking, based on a data warehouse, but on document repositories indexed with Lucene.

However, despite the increasing movement toward data repurposing, many factors might still limit the adoption of CDWs: limited graphical interfaces, the use of query languages (e.g. SQL), the need for prior knowledge of ontologies, and so forth. In addition, the exploitation of the data often requires advanced abilities with statistical or analytical software.

A Rare Disease Context

The *Necker Enfants Malades Hospital* (Necker Children's Hospital) is a national reference center for rare and undiagnosed diseases. The hospital hosts the *Imagine* Institute, a research institute focused on genetic diseases. As part of its data science platform, the *Imagine* Institute has been developing since 2015 a document-oriented data warehouse (Dr Warehouse®), which is intended to meet three specific usage scenarios: clinical research, detection of hypotheses or data mining, and translational research.

Rare diseases exhibit challenging characteristics namely (i) the data are particularly scattered in several databases focusing on translational research (in addition to the classic EHR), (ii) text is of prime importance to describe the clinical presentation of the patient, especially for undiagnosed patients (iii) historical clinical databases are inestimable sources of knowledge, especially when a patient has complex or atypical clinical signs. In such cases, literature may be insufficient and only the comparison with similar patients makes it possible to understand the mechanisms of the disease and to guide the clinician in the therapeutic attitude [29].

A data warehouse oriented toward textual data is an effective way of making the data accessible and preserving it without degrading its quality. Such data warehouse helps to ensure that the system is vendor-neutral and can serve many translational research purposes.

In this article, we introduce Dr.Warehouse (DrWH) an open-source document-oriented data warehouse focused toward clinicians. We developed the data model, the user interface and the functionalities to take advantage of narrative reports. In the following

section, we describe the key technical features, as well as clinical and research functionalities. And we present the first feedback from early adopters and discuss our findings.

2. Method

2.1. Technical

2.1.1. Database model

Dr Warehouse is a document-oriented warehouse. Therefore, we placed the narrative reports in the middle of the database model (instead of the facts in classical CDW architecture). Figure 1 shows a simplified schema of the database organization. The table DWH_DOCUMENT is at the center of the star schema. This table contains the narrative report in ASCII format (PDF, Word and other formats of documents are converted in ASCII). Related to the table DWH_DOCUMENT, we added the patient table (DWH_PATIENT) containing demographics data and the fact table DWH_DATA containing coded data (e.g. biological tests results, diagnosis codes). We stored the movements of the patients (hospital stays, outpatient encounters) in the DWH_PATIENT_STAY table. DWH_DATA is linked to the thesaurus table DWH_THESAURUS_DATA. We also take into account family relations between patients (DWH_PATIENT_REL).

2.1.2. Search engine

Displayed and searchable documents. The development of Dr Warehouse was organized around two principles: (i) a document has to be displayed to the user as close as possible to its original format; (ii) noise in free text search due to negations or family history sections in documents must be reduced as much as possible. Consequently, the system manages two versions of a single document: the displayed document and the searchable document. The process is described Figure 2.

The displayed document is converted to ASCII format and saved in the DWH_DOCUMENT table. The searchable document is stored in a table dedicated to free-text search (DWH_TEXT) along with transformed version of the documents optimized for the search engine.

Dealing with negation and context. A process based on regular expressions splits the searchable texts into sentences and propositions. The process detects and regroups sentences related to the patient or to their family history, as well as propositions expressing or not negations. For each group of propositions, a triplet (propositions; context [patient, family history]; certainty [not negation, negation]) is stored in the DWH_TEXT table. The details of the processing are described in [30].

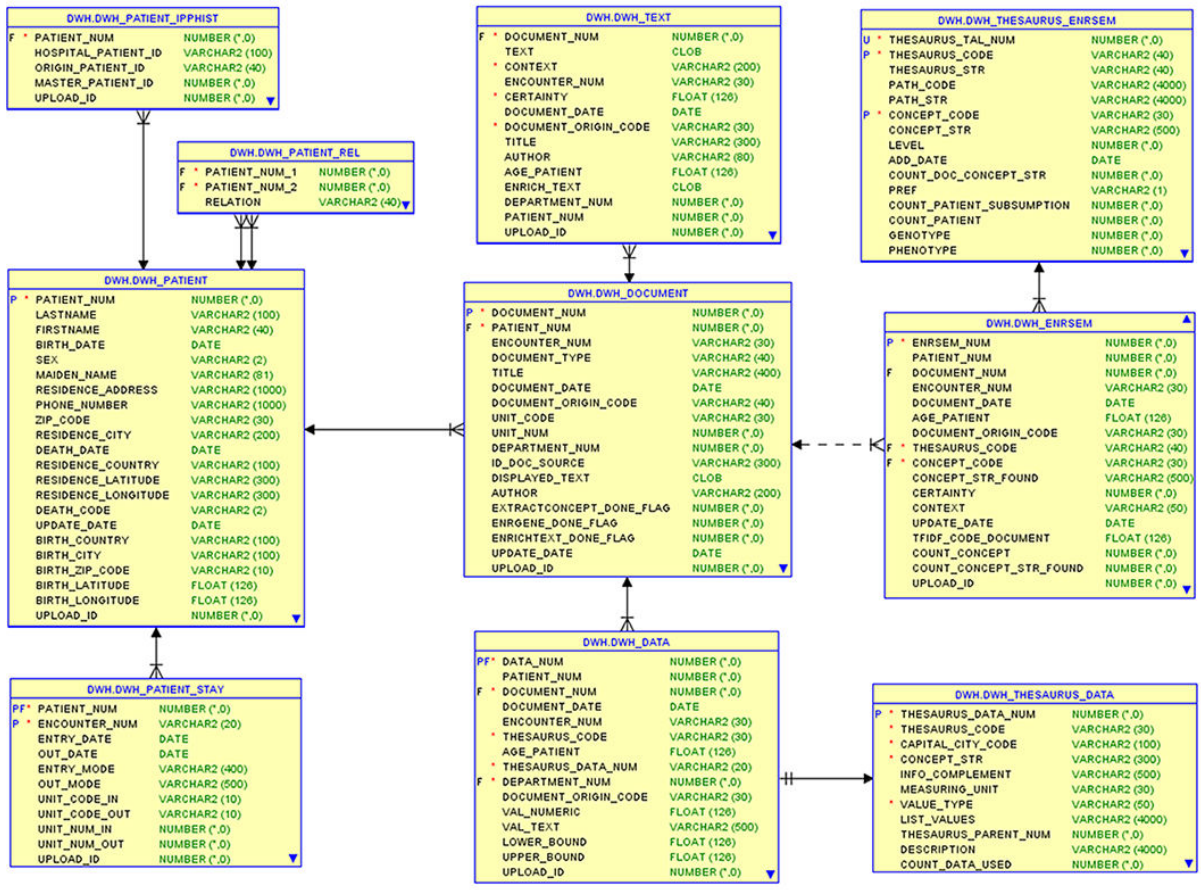


Figure 1: Dr warehouse reduced relational model

We leveraged the Oracle Text Module [31] to index all the sentences of the searchable documents.

Note that Dr Warehouse can be used in two modes: anonymized or with the full identity of the patients. However, the search engine does not allow queries with patient names, as a dedicated function is available for that specific use case. To prevent the search by name in the full text, the searchable documents are anonymized as much as possible: the first and last names, addresses and phone numbers are removed from the documents.

Structured data. We store all structured data in the fact table DWH_DATA (similar to the i2b2 OBSERVATION_FACT table [32]). All structured data are linked to a document with a mandatory foreign key. If no document is associated with the structured data in the original EHR, a document is created, and populated with the concatenation of all coded data from a given source. A document is created per patient, date and hospital departments. For example, all the ICD10 codes and labels of the patient 1, on 2005-06-25 from the cardiology department will be grouped in a document, codes from another department, or on another day would be stored in a distinct document. Thus, the user is able to search this document by using a full-text query and/or a coded query because both are stored in the data warehouse.

2.1.3. Semantic annotation and enrichment.

Recognizing UMLS® phenotype concepts. We leverage the Unified Medical Language System Metathesaurus® (UMLS®) [33] to allow query expansions (more specifically adding synonyms and subsumption relations). We created a phenotype subset of the UMLS by selecting all the terms (including all synonyms) in the language of the narrative reports (e.g. LAT='FRE' in MRCONSO table) and categorized in one of 10 semantic types related to phenotype information: 'Sign or Symptom', 'Disease or Syndrome', 'Finding', 'Pathologic Function', 'Congenital Abnormality', 'Physiologic Function', 'Anatomical Abnormality', 'Neoplastic Process', 'Acquired Abnormality' and 'Mental or Behavioral Dysfunction'. We use this subset with an exact match strategy to recognize UMLS concepts on the searchable documents. Prior to matching, we perform a shallow normalization (inflection, termination, special characters). We also extended the list of stop words using a prefix strategy (e.g. the word following the prefix "Dr." is ignored). In the remaining of this article, we will refer as phenotype all the concepts of this subset.

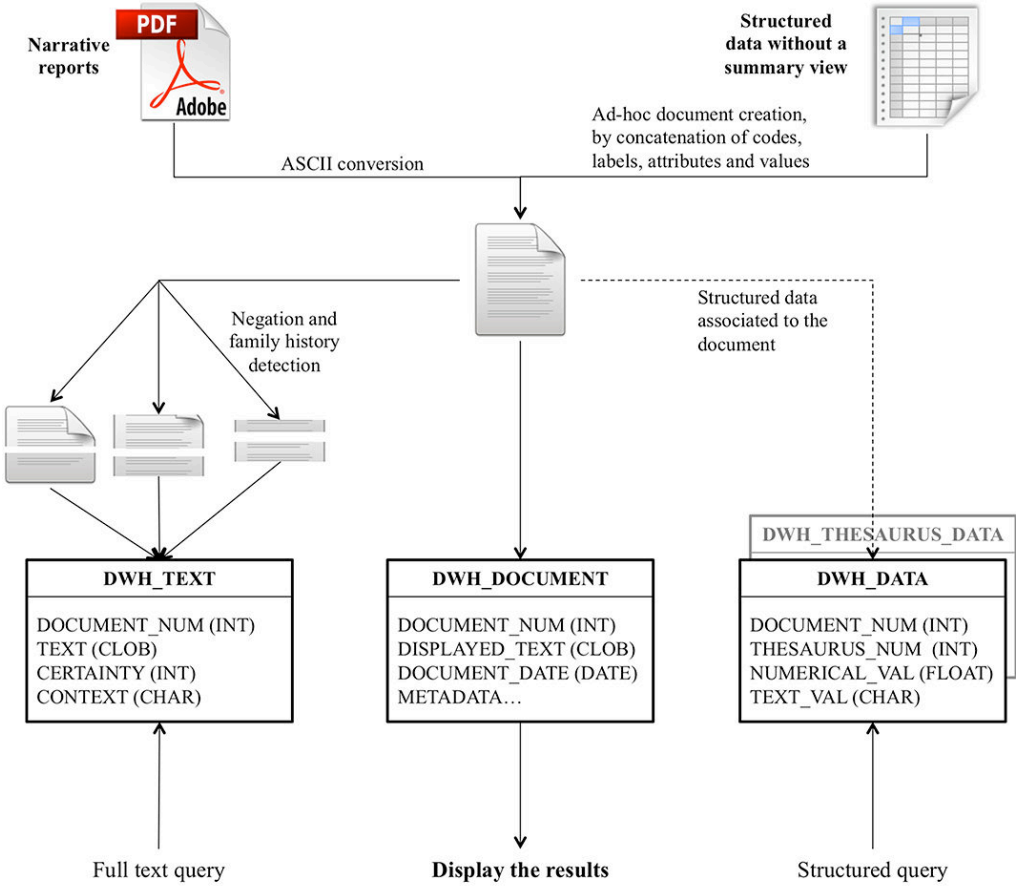


Figure 2: The ETL process from a document to the DrWH tables. This figure illustrates two use cases, (i) integration of a narrative report associated with possibly structured data, (ii) integration of structured data without a summary view (ICD codes). Both types of document are available for full text query.

Recognizing genes. We created a gene subset of the UMLS by selecting all the terms in the UMLS categorized in "Gene or Genome" semantic type. We kept gene names having more than 3 characters and less than 55 characters [34] from the thesauri 'OMIM' or 'HUGO Gene Nomenclature'. We applied a pre-processing step to take into account the case sensitivity to limit the number of false positive matches due to homonyms [35]. We processed as case sensitive the terms without numbers (e.g. "NEMO" UMLS: C1416380). We processed as case insensitive the terms with a number (e.g. "MECP2" UMLS: C1417098). Then we extracted the gene names contained in documents using a case-sensitive (resp. insensitive) exact match.
Note that these data recognition strategies (gene and phenotype) are independent of Dr WH itself. The administrator has to ensure that the ETL process correctly loads the data into the appropriate tables and format.

Storing concepts. We stored all the UMLS concepts identified during the semantic annotation for two use cases: (i) for query expansion: We added a new indexed column to the DWH_TEXT table in which we concatenated the original clinical text (part of speech per certainty and context) , the concepts extracted, their UMLS synonyms and their UMLS ancestors; (ii) For phenotyping: we stored for each concept the triplet (concept UMLS identifier (CUI); context [patient, family history]; certainty [not negated, negated]) in the table DWH_ENRSEM. In addition, we stored the matching term associated to the CUI in the CONCEPT_STR_FOUND column. For example, in the sentences "He presents an Oculocerebrorenal syndrome" and "He was diagnosed with Lowe syndrome", we stored "C0028860" and "Oculocerebrorenal syndrome", "C0028860" and "Lowe syndrome" respectively. We determined the frequency of terms in narrative documents to flag the "preferred terms" according to the perspective of the clinicians. This process is described in Figure 3.

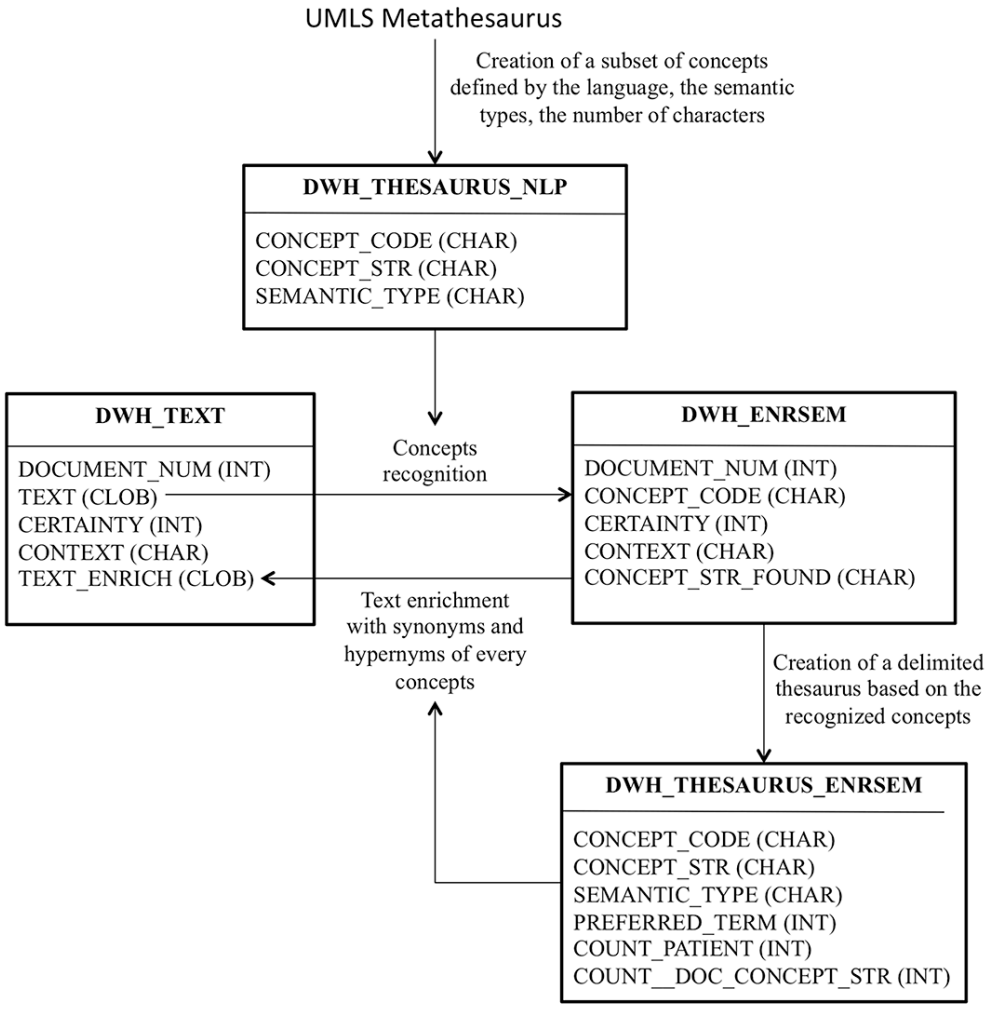


Figure 3: The process of concepts recognition based on a subset of the UMLS metathesaurus. The recognized concepts are used to create a enrich text which can be used to enhance the sensitivity of the search engine

2.1.4. Data access optimization

We chose to denormalize the database schema to optimize the performance of the CDW [36]. We duplicated columns used by the search engine to reduce the number of joins in SQL queries. The metadata describing a document (patient Id, Encounter Id, Patient age, Date, Title, Hospital department etc.) are stored three times in the DWH_DOCUMENT, DWH_TEXT and DWH_DATA tables.

We pre-calculated aggregated information to accelerate the user experience. For example, the number of patients and documents per concept are stored in the DWH_THESAURUS_ENRSEM table. These frequencies are used in the translational features to calculate the TF-IDF scores.

The terminological enrichment requires exploiting the hierarchy of the thesaurus. More specifically, we enrich the original text with all the synonyms and hypernyms corresponding to the UMLS concepts found in the document.

We manage the complexity of the hierarchy by pre-computing the transitive closure along with the distance between concepts (more than one distance can be associated with a pair of concepts). We built the table DWH_THESAURUS_ENRSEM_GRAPH in which we store each concept of the DWH_THESAURUS_ENRSEM. We built this table by using the MRHIER table of the UMLS Metathesaurus [37].

2.1.5. Implementation

. We developed Dr Warehouse using the PHP 5.3 language (note that the source code is compatible with PHP7) and Oracle® v.11g for the processing (compatible with Oracle® v12). We leveraged HTML4 and jQuery v1.11, GraphViz [38], the HighChart library for the visualization of the aggregated results and the Google Map API for the patients' geographical distribution.

The source code is available at this github link https://github.com/imagine-bdd/DRWH .

A demonstration is available at https://imagine-plateforme-bdd.fr/dwh_pubmed/ . For this demonstration, the data warehouse has been fed with PubMed abstracts.

2.2. Clinician oriented interface

Whereas the back end of the data warehouse is similar to other approaches, we have concentrated the development of DrWH around several user-friendly graphical interfaces.

Our goal was to build a simple and efficient software application for daily usage and the enabling secondary use of clinical data. DrWH was designed in collaboration with clinicians and clinical research assistants to fit as closely as possible to their needs.

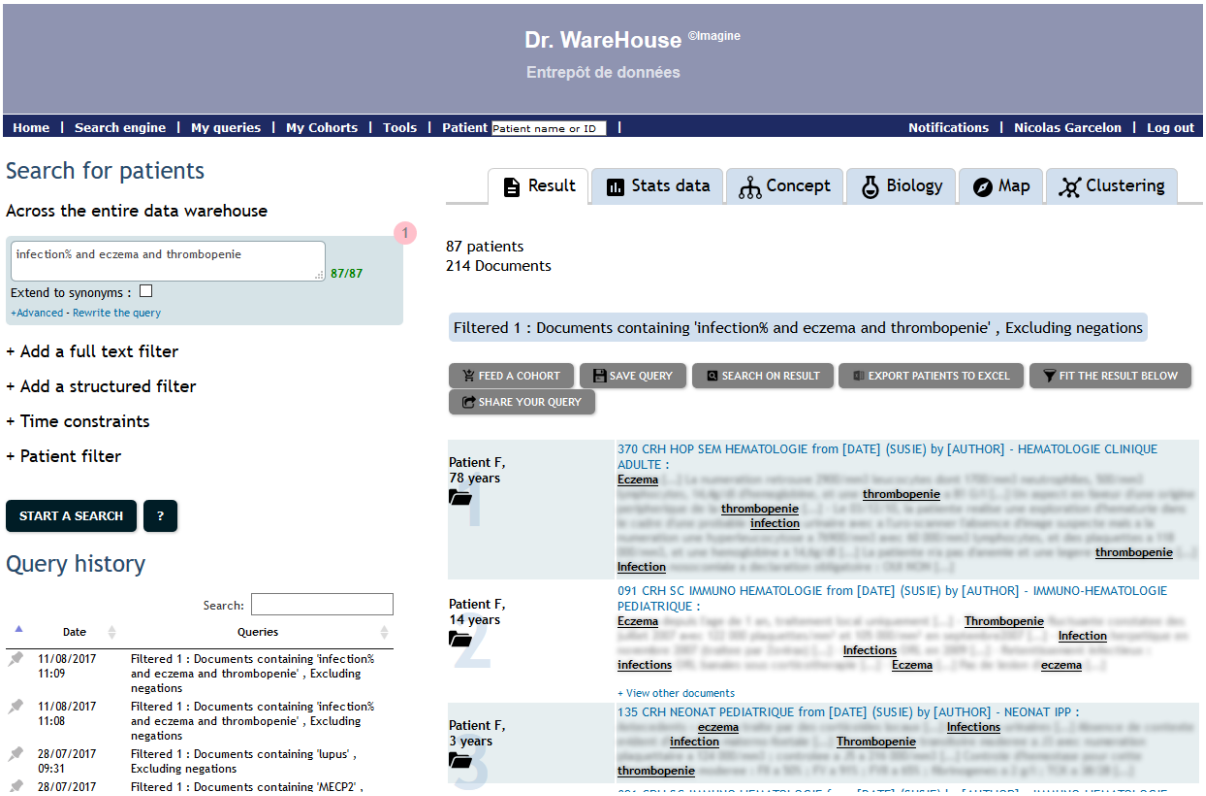


Figure 4: Dr Warehouse search engine interface. One atomic search has been made: "infection% and eczema and thrombocytopenia". The search engine found 87 patients and 214 documents in which the 3 required words were found. The first tab "result" is open and displays the documents found grouped by patient. The capture has been anonymized for the illustration. The system highlights the search words. To visualize the document, the user clicks on the title.

2.2.1. Clinical research and studies / cohort

Searching documents and cases. The main feature of the data warehouse is its search functionality. Users can query Dr Warehouse in a "Google-like" way. The user enters one or more words (symptoms, disease names, gene names...), and the system returns a set of patients with relevant documents. The user can also build queries using coded data (e.g. biological results) and easily combine free text and coded queries by creating several atomic queries. In real time, the system calculates the number of patients found for each atomic query, but waits for the confirmation of the user to compute the intersection of atomic queries and display the list of patients matching all of the atomic queries. Note that by default the system searches in positive statement (i.e. not negated and regarding the patients themselves, not family history, see section 2.1.2).

Visualizing matches. The text-based search engine may display false positive patients; consequently we display on the graphical user interface (GUI) a portion of the text containing the researched terms (even if the text found is a synonym or a hyponymy of the query). The users can easily verify the relevance of the results and may modify their query if needed or exclude the patient from the result set (Figure 4).

Advanced querying and query expansion. The search engine allows advanced search. Users can build sets of constraints, including time constraints (length of the follow-up or time dependencies between two atomic queries) and demographic constraints (age and sex of the patients). The user can also enable an automated query expansion that leverages the synonyms and subsumption relations of the UMLS. By default, a subatomic query is considered as an inclusion criterion, the user can modify this as an exclusion criterion.

Simplifying clinical research and improving users' experience. Several features have been developed to improve the clinicians' experience, including:

- *Misspelling correction*: When fewer than 10 patients are identified by an atomic query, the system displays the five most similar queries. Similar queries are suggested based on the Levenshtein distance with UMLS corpus and successful users' queries stored in a dedicated table.
- *Cohort and patient recruitment*: The result tab of the GUI allows the user to create a cohort of patients and include or exclude identified patients in this cohort. Several users can share the same cohort to enable collaborative inclusion.
- *Automated alert*: The users can save queries. The queries will be executed automatically every month and Dr Warehouse will alert the user if new patients matched their query. A dedicated interface allows the new patients to be managed.
- *Recursion of the search engine*: A user can choose to run the search engine (i) on the entire data warehouse, (ii) on a datamart built from the results of a prior query, or (iii) on a datamart built from a cohort of patients.
- *Shared queries*: Users can easily share a query with each other. This feature can also be used by a hotline to teach or assist users in the construction of complex queries. A simple click on a shared query fills the search engine form. The queries are stored in XML format and could be shared among several DrWH instances.

2.2.2. Phenotype Mining

In addition to the functionalities previously described (2.2.1) displaying the list of patients and documents found, the search engine displays several features presenting aggregated data related to the identified population.

Stats data tab. The GUI provides a demographic description of the population: sex ratio, age structure, death rate, number of patients per year and number of patients per hospital department.

This tab displays also aggregated care path. The GUI shows the flows of the patients through the hospital departments. The edges thickness is calculated according to the number of patients. Figure 5 represents the aggregated care path of patients found with the query "lupus" in Necker Hospital DrWH. This GraphViz representation allows to visualize none linear care path common in the context of rare diseases. We also display a Sankey representation allowing an interactive linear representation of the care patt.

Concepts tab. The "**concepts tab**" displays the list of UMLS concepts extracted from the narrative reports and the number of occurrences (i) at the document level (concepts from the documents matching the query) or (ii) at the patient level (all the concepts of patient matching the query). The first case is especially relevant for acute diseases or acute episodes, and the latter can be useful for example in the context of chronic or rare diseases.

The ULMS concepts are displayed in a table and are sortable by common metrics (frequency and TF-IDF [39]) as well as two metrics dedicated to the exploration of phenotype associations. We display the terms flagged as "preferred terms" during the UMLS annotation process to remain as close as possible to the clinicians' vocabulary. The extraction of UMLS concepts can help address two use cases:

- Phenotype description: in this case, the query is a syndrome or a gene name, the table sorts the most frequent phenotypes associated with the query. (e.g. Figure 6)
- Diagnosis hypothesis: the query is an association of clinical signs, the table sorts the diagnosis strongly associated with the query. (e.g. Figure 7)
- *Map tab*. The "**map tab**" displays the position of the patients on Google Map by using their zip code.

This representation may be useful to explore epidemiological studies or hospital reputation.

- *Biology tab*. The "**lab tab**" displays the aggregated biological results of the patients in a sortable table (minimum, maximum, average, number of values over the upper bound, number of values under the lower bound, and so forth).

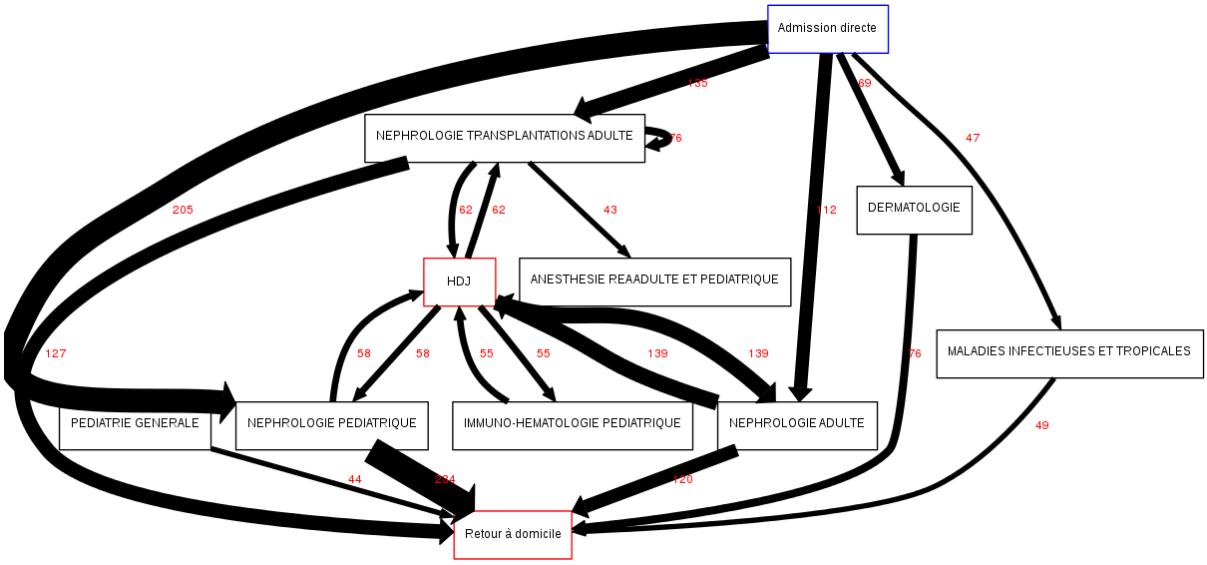


Figure 5: Care path of patients found with the query "lupus" in Necker Hospital DrWH.

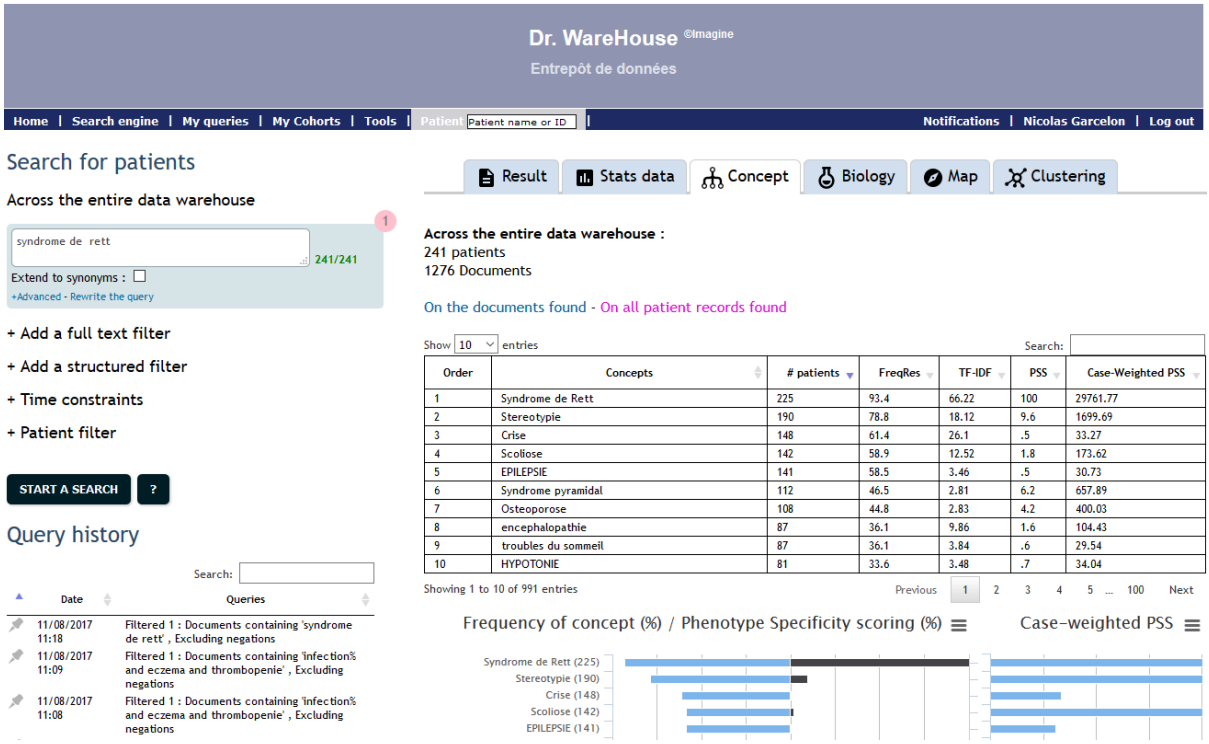


Figure 6: The "concept tab" displays the UMLS concepts recognized in the patient's records found with the query "Rett syndrome". The concepts are ranked by frequency. In this example, as found in the literature, the main phenotypes associated to Rett syndrome are: Stereotypy, Crisis, Scoliosis, Seizure, Pyramidal syndrome, Osteoporosis, Encephalopathy, Sleep disorder, Hypotonia.

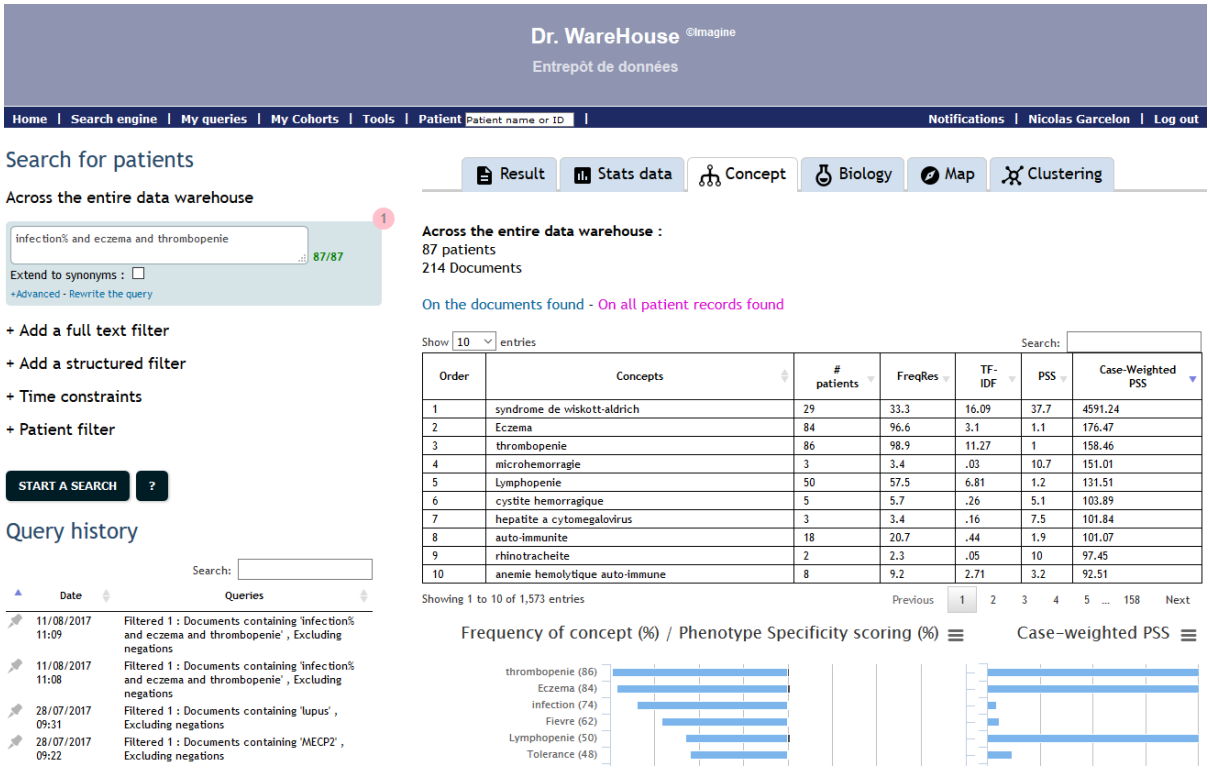


Figure 7: The "concept tab" displays the UMLS concepts recognized in the patients records found with the query "infection% and eczema and thrombocytopenia". The concepts are ranked by TF-IDF score. In this example, "Wiskott Aldrich syndrome" is the most specific concept associated with the searched phenotype association.$

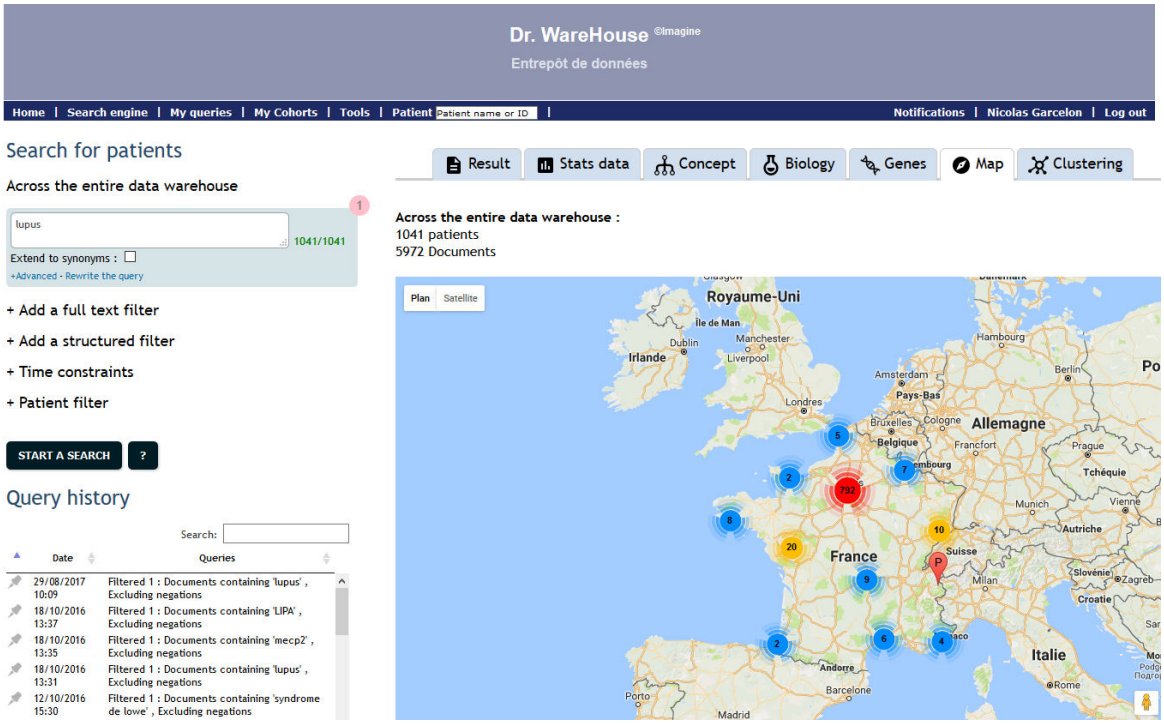


Figure 8: The "map tab" displays a Google Map of the patients found with the query "lupus".

2.2.3. Patient centric / Translational research

The main feature of many CDWs is the identification of cohorts. The patient-centric view provides the clinician with multiple ways to explore the medical history of a single patient:

- A search engine restricted to the patient's documents. It is based on the same technology as the global search engine.
- A searchable timeline, which allows the user to explore the patient history with the view of the

duration between the medical events. We used Simile Timeline developed by the MIT [40]. A search engine is also integrated and highlights the event containing the documents with the searched terms.
- The family tree of the patients, specifically relevant in the context of rare diseases and genetic diseases.
- The care path of the patient provides the clinician with a global view of his journey.
- The cohorts in which the patient is included or excluded.
- The list of UMLS concepts recognized in the narrative reports. The concepts are ranked by their frequency thus the user has access to a medical profile of the patient.
- Advanced "similar patients" functionalities using the similarity algorithm described in Garcelon *et al.* 2017 [41].

2.2.4. Administration tools

The administrator of Dr Warehouse can fine-tune users' profiles through a dedicated administration GUI. Each feature can be activated or not according to a user's profile. The user may have access to the entire data warehouse or only to patients who have come to their hospital department at least once. The administrator may also allow access to all or some of the integrated data sources (for example we can exclude access to the genetic data from a user's profile). And finally, the administrator can define profile with de-identified or nominative access to the data.

All the queries and clicks are traced with the user's id and the access date of each data.

2.2.5. Persistent environment

The user's desktop is persistent from a session to a new one. The user may use DrWH as a dedicated environment for his research works. We designed several GUI to allow him to manage his data (cohorts, queries, notifications, and so forth...).

3. Results

3.1. Data

We implemented Dr warehouse at *Necker Enfants Malades Hospital* (Necker hospital). This hospital is a 600 beds University hospital located in Paris, France. Necker Hospital is a major referral center for rare diseases and for complex pathologies. The hospital Necker-Enfants Malades provides services well beyond Paris and its area: more than 20% of the patients come from other provinces in France or from aboard [42].

Up to July 2017, the data warehouse contained approximately 480,000 patients, 4 million clinical documents, 34 million structured data items (biological results, ICD10 codes) and 15.5 million recognized UMLS concepts. We integrated 21 sources of data including the hospital EHR database, the biological result database, the imaging database, the DRG (ICD10), specific databases such as the foetopathology database, the antenatal ultrasound database and several deprecated databases. The oldest data goes back to 1996. Table 1 describes the distribution of recognized concepts and documents per patient.

3.2. Usage statistics

Dr Warehouse was partially deployed in Necker Hospital in January 2017. The software was presented during the clinical staff meetings. We did not organize any training sessions. We evaluated the use of Dr Warehouse for a 6 months period (from January until June 2017). The most represented departments are Pediatric Nephrology, Metabolism and pediatric neurology, Adult clinical hematology, Pediatric hematology immunology, Pediatric Cardiology, Pediatric otorhinolaryngology, Pediatric radiology.

Clinical research and studies / cohorts. 122 distinct users connected 1946 times to Dr Warehouse (239 in January, 374 in June). They executed 2837 queries in the search engine. They saved 173 queries for automated monthly alert. During this period they created 131 cohorts in which they included 36,632 patients and excluded 4,267 patients. The number of screened patients is detailed in Table 2.

Patient centric view. The users visualized a total of 13,366 patients' records and executed 3,999 queries inside a patient record to find a specific document. They opened 34,733 clinical reports. And they used 111 times the similarity feature.

3.3. End user satisfaction survey

During the presentation in clinical staff meetings, the reactions were highly positive.

We conducted a survey in August 2017 using the same questionnaire as the Emerse survey [17]. We received responses from 59 users. The users' profiles are 81% clinicians with research activities, 14% are clinical research assistants and 5% are data scientists.

Table 3 shows that the vast majority of the users is satisfied and thinks the system reduced their time-consuming tasks (96.6%). Only 8.5% of the users report some difficulties to find accurate patients with the search engine.

Table 1: Distribution of the recognized concepts and documents per patient in the data warehouse

	Average	Standard Deviation	Median	Min	Max
# documents per patient	7.93	25.46	3	1	1,654
# concepts per patient	51.66	148.71	16	0	7,288
# distinct concepts per patient	16.43	24	9	0	568
# concepts per document	10.27	14.13	6	0	1,433
# distinct concepts per document	8.36	10.48	5	0	322

Table 2: Median number of patients screened by users' profile

User profiles:	ARC	Clinician	Data scientist
Median number of screened patients	978	74	1,996
Median number of included patients	787	46	974
Median number of created cohorts	8	2	8

Table 3: End user satisfaction survey results

	Does not meet (%)	Meets or exceeds (%)	Does not apply (%)
1. Enables effective searching	1.7	96.6	1.7
2. Solves my problem or facilitates the tasks I face	3.4	94.9	1.7
3. Saves overall time and effort	0.0	96.6	3.4
4. Allows me to get accurate answers	8.5	78.0	13.5
5. Helps me find data I might have otherwise missed or overlooked	1.7	86.4	11.9
6. Has expanded my ability to conduct chart reviews to areas previously impossible with manual review	3.4	79.7	16.9

4. Discussion

We built the data warehouse model and conceived the algorithms in a way that DrWH can work without any coded data and without any NLP algorithm. Dr WH can be deployed rapidly with basic functionalities regardless of the language of the narrative reports. With additional resources, NLP algorithms can be applied during the ETL process to enhance the search engine, and to recognize medical concepts for the phenotypic and similarity features.

Indeed, the terminological alignment of coded data and the NLP are the most expensive parts of building a CDW, especially for languages other than English. Many NLP algorithms have been developed for the English language including cTakes [43], Sophia [44], Metamap [45], Negex[46], Medex [47]. Conversely, only few NLP algorithms exist for others languages [48]. In addition, a limited number of annotated clinical corpora are available for non-English languages. For the French language, we have identified QUAERO [49], MANTRA GSC [50] and MERLOT [51]. But only the latter contains real clinical reports. However MERLOT is not publicly available due to privacy concern.

4.1. Comparison to other CDWs

Most of the CDWs are dedicated to the exploitation of coded data. Users do not have access to an interface dedicated to the validation of identified cases (cases found are considered as true positives). Moreover, this type of architecture implies a good knowledge of the structure of the data, which is rarely the case with clinicians. Jannot *et al.* [52] highlighted the necessity for the clinicians to work with the medical informatics and bioinformatics department in order to build efficient queries.

There is still a huge amount of new knowledge to discover from the clinical narratives [53,54], especially in the context of rare diseases. DrWH is natively developed to handle “dirty data” in the form of clinical narratives. Such textual data require a proper GUI for efficient validation and exploration.

Emerse [12] was created with the same spirit to take advantage of textual data. However they do not propose automated solutions to exclude false positive due to negation or family history. The DrWH database schema enables multiple contexts in addition to family history (patient history, reason for

admission, conclusion, results, etc.). Thereby, the users can focus a query on a specific part of the text.
Another originality of DrWH lies in the patient-centric view and the translational tools integrated in the GUI (similarity, high throughput phenotyping). An important effort has been made to optimize processing times to allow complex computations in online sessions.
We also developed an environment dedicated to the day-to-day translational research activity of clinicians. DrWH enhanced collaborative works around cohorts' constitution and data mining. Our solution aims at providing user-friendly interfaces and interactions with clinicians and researchers to enable mining of patient data and phenotype exploration.
Another category of warehouses has emerged dedicated to the integration of clinical and omics data [55]. DrWH does not aim at the integration of omics data. But structured omics information can be stored as coded data. Nevertheless, it would be simple to export cohorts of patients from DrWH to omics-oriented platforms.

4.2. The use of Dr Warehouse at the Necker Hospital

The number of queries per month increased by 56% from January to June 2017. The queries made by the users are most of the time simple queries with one atomic full text query. However the complexity of the queries increased along with the user experience. The system was extensively used for clinical research: clinical research assistants used DrWH to screen a median of 978 patients (787 patients included in the studies).
The exploration of patients' records is widely used (over 13,000 records viewed in 6 months). The "inside patient's record" search engine is appreciated by the users (about 4,000 queries). By analyzing the users queries, this "inside patient's record" search engine is mostly used to find specific events (e.g. biopsy, MRI, scintigraphy), treatments (e.g. cystagon, rituximab, Bactrim) or biological results.
Besides clinical research, we expect that clinicians will use more available features (phenotype associations, patient similarity) as we plan to hold training sessions in the second semester of 2017.
The clinicians now have to deal with a new way to use the patient records for research in a data-driven approach. This is illustrated by the 16.9 % of "does not apply" for the question 6 of the survey ("*Has expanded my ability to conduct chart reviews to areas previously impossible with manual review*"). Users have occasionally highlighted this concern during Dr. Warehouse's presentation to the clinical staff meetings. Five users do not consider that DrWH help them obtain accurate answers. While DrWH can still be improved, most of the false positive patients could be managed by the use of exclusion criteria. This highlights the need for proper training sessions for advanced functionalities.
Otherwise, feedback from the users was globally extremely positive and DrWH saved time for research (96.6% answered that DrWH enabled effective searching).

4.3. Ethics and security

Grande *et al.*[56] showed that the main concern of patients was the specific purpose of using their data. The second concern was the user of this information. The sensitivity of the data was not a significant issue. Another study showed that most of patients agree to let their data used for clinical research [57].
DrWH was designed to be used in a confined environment (i.e. inside the hospital). The software stores personal narrative data, which are highly sensitive. We developed DrWH for the use of clinicians in charge of the patients. DrWH allows user profiles to be managed with specific rights in compliance with the hospital and the ethics committee policies.
We strongly suggest that DrWH must be hosted on a server inside the hospital network to benefit from the highest security level. Clinicians are responsible for the use of the data warehouse. All metadata associated to the log files can be mined for ethical and security check.

4.4. Dissemination

DrWH is installed since June 2017 at the European Hospital George Pompidou (HEGP) in Paris. HEGP was the first hospital in France to use i2b2. HEGP data were exported from the i2b2 instance in which coded data and clinical reports were stored to DrWH. HEGP now has two data warehouses for different use cases: DrWH is dedicated for use by clinicians, i2b2 for data scientists.
The Foch hospital (a 600 beds French private hospital, located in Suresnes, in the suburbs of Paris) is in the process of setting up DrWH by the end of 2017.

5. Conclusion

We have developed Dr Warehouse to provide clinicians with an ergonomic software application to screen patients by searching among millions of narrative clinical notes. The main functionalities include free text search, cohort recruitment, high throughput phenotyping and patient similarity.
The system has been used in production for several months at Necker Hospital with a 96.6% satisfaction rate. DrWH is open source and available online (https://github.com/imagine-bdd/DRWH). A demonstration based on PubMed abstracts is available at https://imagine-plateforme-bdd.fr/dwh_pubmed/.

6. Acknowledgement

Bastien Rance is supported in part by the SIRIC CARPEM cancer integrated research program.

We thank Guillaume Huart, Alain Fischer, Christophe Nicolaï, Laure Boquet and Stanislas Lyonnet, for their support in this project. We thank the scientific committee of Dr WH for their help. We thank Pauline Touche for her constructive comments.

References

[1] S.J. Hebbring, M. Rastegar-Mojarad, Z. Ye, J. Mayer, C. Jacobson, S. Lin, Application of clinical text data for phenome-wide association studies (PheWASs), Bioinforma. Oxf. Engl. (2015). doi:10.1093/bioinformatics/btv076.
[2] B. Namjou, K. Marsolo, R.J. Caroll, J.C. Denny, M.D. Ritchie, S.S. Verma, T. Lingren, A. Porollo, B.L. Cobb, C. Perry, L.C. Kottyan, M.E. Rothenberg, S.D. Thompson, I.A. Holm, I.S. Kohane, J.B. Harley, Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis, Front. Genet. 5 (2014) 401. doi:10.3389/fgene.2014.00401.
[3] J.C. Denny, M.D. Ritchie, M.A. Basford, J.M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D.R. Masys, D.M. Roden, D.C. Crawford, PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations, Bioinforma. Oxf. Engl. 26 (2010) 1205–1210. doi:10.1093/bioinformatics/btq126.
[4] A. Neuraz, L. Chouchana, G. Malamut, C. Le Beller, D. Roche, P. Beaune, P. Degoulet, A. Burgun, M.-A. Loriot, P. Avillach, Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics, PLoS Comput. Biol. 9 (2013) e1003405. doi:10.1371/journal.pcbi.1003405.
[5] M. Orsini, A. Travaglione, E. Capobianco, Warehousing re-annotated cancer genes for biomarker meta-analysis, Comput. Methods Programs Biomed. 111 (2013) 166–180. doi:10.1016/j.cmpb.2013.03.010.
[6] E.S. Chen, I.N. Sarkar, Mining the electronic health record for disease knowledge, Methods Mol. Biol. Clifton NJ. 1159 (2014) 269–286. doi:10.1007/978-1-4939-0709-0_15.
[7] M.D. Krasowski, A. Schriever, G. Mathur, J.L. Blau, S.L. Stauffer, B.A. Ford, Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research, J. Pathol. Inform. 6 (2015) 45. doi:10.4103/2153-3539.161615.
[8] C. Shivade, P. Raghavan, E. Fosler-Lussier, P.J. Embi, N. Elhadad, S.B. Johnson, A.M. Lai, A review of approaches to identifying patient phenotype cohorts using electronic health records, J. Am. Med. Inform. Assoc. JAMIA. 21 (2014) 221–230. doi:10.1136/amiajnl-2013-001935.
[9] D.J. Odgers, M. Dumontier, Mining Electronic Health Records using Linked Data, AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci. 2015 (2015) 217–221.
[10] H. Cao, M. Markatou, G.B. Melton, M.F. Chiang, G. Hripcsak, Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics, AMIA. Annu. Symp. Proc. 2005 (2005) 106–110.
[11] P. Raghavan, J.L. Chen, E. Fosler-Lussier, A.M. Lai, How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?, AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci. 2014 (2014) 218–223.
[12] D.A. Hanauer, Q. Mei, J. Law, R. Khanna, K. Zheng, Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE), J. Biomed. Inform. 55 (2015) 290–300. doi:10.1016/j.jbi.2015.05.003.
[13] R. Charon, At the membranes of care: stories in narrative medicine, Acad. Med. J. Assoc. Am. Med. Coll. 87 (2012) 342–347. doi:10.1097/ACM.0b013e3182446fbb.
[14] A. Groß, C. Pruski, E. Rahm, Evolution of biomedical ontologies and mappings: Overview of recent approaches, Comput. Struct. Biotechnol. J. 14 (2016) 333–340. doi:10.1016/j.csbj.2016.08.002.
[15] S.T. Rosenbloom, J.C. Denny, H. Xu, N. Lorenzi, W.W. Stead, K.B. Johnson, Data from clinical notes: a perspective on the tension between structure and flexible documentation, J. Am. Med. Inform. Assoc. JAMIA. 18 (2011) 181–186. doi:10.1136/jamia.2010.007237.
[16] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, H.C. Chueh, Integration of clinical and genetic data in the i2b2 architecture, AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp. 2006 (2006) 1040.
[17] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J. Am. Med. Inform. Assoc. JAMIA. 17 (2010) 124–130. doi:10.1136/jamia.2009.000893.
[18] Observational Health Data Sciences and Informatics. Analytic tools, (n.d.). https://www.ohdsi.org/analytic-tools/ (accessed August 12, 2017).
[19] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers, Stud. Health Technol. Inform. 216 (2015) 574–578.
[20] H.J. Lowe, T.A. Ferris, P.M. Hernandez, S.C. Weber, STRIDE--An integrated standards-based translational research informatics platform, AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp. 2009 (2009) 391–395.
[21] I. Danciu, J.D. Cowan, M. Basford, X. Wang, A. Saip, S. Osgood, J. Shirey-Rice, J. Kirby, P.A. Harris, Secondary use of clinical data: the Vanderbilt approach, J. Biomed. Inform. 52 (2014) 28–35. doi:10.1016/j.jbi.2014.02.003.
[22] C.G. Chute, S.A. Beck, T.B. Fisk, D.N. Mohr, The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data, J. Am. Med. Inform. Assoc. 17 (2010) 131–135. doi:10.1136/jamia.2009.002691.
[23] D.L. Rubin, T.S. Desser, A data warehouse for integrating radiologic and pathologic data, J. Am. Coll. Radiol. JACR. 5 (2008) 210–217. doi:10.1016/j.jacr.2007.09.004.
[24] H. Hu, M. Correll, L. Kvecher, M. Osmond, J. Clark, A. Bekhash, G. Schwab, D. Gao, J. Gao, V. Kubatin, C.D. Shriver, J.A. Hooke, L.G. Maxwell, A.J. Kovatich, J.G. Sheldon, M.N. Liebman, R.J. Mural, DW4TR: A Data Warehouse for Translational Research, J. Biomed. Inform. 44 (2011) 1004–1019. doi:10.1016/j.jbi.2011.08.003.
[25] M. Puppala, T. He, S. Chen, R. Ogunti, X. Yu, F. Li, R. Jackson, S.T.C. Wong, METEOR: An Enterprise Health Informatics Environment to Support Evidence-Based Medicine, IEEE Trans. Biomed. Eng. 62 (2015) 2776–2786. doi:10.1109/TBME.2015.2450181.
[26] P. Heudel, A. Livartowski, P. Arveux, E. Willm, C. Jamain, [The ConSoRe project supports the implementation of big data in oncology], Bull. Cancer (Paris). 103 (2016) 949–950. doi:10.1016/j.bulcan.2016.10.001.
[27] K.U. Kortüm, M. Müller, C. Kern, A. Babenko, W.J. Mayer, A. Kampik, T.C. Kreutzer, S. Priglinger, C.

Hirneiss, Using electronic health records to build an ophthalmological data warehouse and visualize patients' data, Am. J. Ophthalmol. (2017). doi:10.1016/j.ajo.2017.03.026.
[28] D.A. Hanauer, EMERSE: The Electronic Medical Record Search Engine, AMIA. Annu. Symp. Proc. 2006 (2006) 941.
[29] J. Frankovich, C.A. Longhurst, S.M. Sutherland, Evidence-based medicine in the EMR era, N. Engl. J. Med. 365 (2011) 1758–1759. doi:10.1056/NEJMp1108726.
[30] N. Garcelon, A. Neuraz, V. Benoit, R. Salomon, A. Burgun, Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse, J. Am. Med. Inform. Assoc. JAMIA. (2016). doi:10.1093/jamia/ocw144.
[31] Oracle Text, (n.d.). http://www.oracle.com/technetwork/testcontent/index-098492.html (accessed April 6, 2017).
[32] Observations » Informatics for Integrating Biology & the Bedside (i2b2) | Boston University, (n.d.). http://www.bu.edu/i2b2/mhdr/mhdr-data/observation_fact/ (accessed August 29, 2017).
[33] D.A. Lindberg, B.L. Humphreys, A.T. McCray, The Unified Medical Language System, Methods Inf. Med. 32 (1993) 281–291.
[34] S.T. Wu, H. Liu, D. Li, C. Tao, M.A. Musen, C.G. Chute, N.H. Shah, Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis, J. Am. Med. Inform. Assoc. JAMIA. 19 (2012) e149–e156. doi:10.1136/amiajnl-2011-000744.
[35] T. Groza, K. Verspoor, Assessing the Impact of Case Sensitivity and Term Information Gain on Biomedical Concept Recognition, PLoS ONE. 10 (2015). doi:10.1371/journal.pone.0119091.
[36] S.K. Shin, G.L. Sanders, Denormalization strategies for data retrieval from data warehouses, Decis. Support Syst. 42 (2006) 267–282. doi:10.1016/j.dss.2004.12.004.
[37] B.T. McInnes, T. Pedersen, S.V.S. Pakhomov, UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity, AMIA. Annu. Symp. Proc. 2009 (2009) 431–435.
[38] Graphviz | Graphviz - Graph Visualization Software, (n.d.). http://www.graphviz.org/ (accessed September 16, 2017).
[39] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1972) 11–21.
[40] SIMILE Widgets | Timeline, (n.d.). http://simile-widgets.org/timeline/ (accessed September 16, 2017).
[41] N. Garcelon, A. Neuraz, V. Benoit, R. Salomon, S. Kracker, F. Suarez, N. Bahi-Buisson, S. Hadj-Rabia, A. Fischer, A. Munnich, A. Burgun, Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack, J. Biomed. Inform. 73 (2017) 51–61. doi:10.1016/j.jbi.2017.07.016.
[42] Introducing the Necker-Enfants Malades hospital, Hôp. Necker-Enfants Mal. (2015). http://hopital-necker.aphp.fr/introducing-necker-enfants-malades-hospital/ (accessed September 20, 2017).
[43] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. JAMIA. 17 (2010) 507–513. doi:10.1136/jamia.2009.001560.
[44] G. Divita, Q.T. Zeng, A.V. Gundlapalli, S. Duvall, J. Nebeker, M.H. Samore, Sophia: A Expedient UMLS Concept Extraction Annotator, AMIA. Annu. Symp. Proc. 2014 (2014) 467–476.
[45] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program., Proc. AMIA Symp. (2001) 17.
[46] W.W. Chapman, D. Hillert, S. Velupillai, M. Kvist, M. Skeppstedt, B.E. Chapman, M. Conway, M. Tharp, D.L. Mowery, L. Deleger, Extending the NegEx lexicon for multiple languages, Stud. Health Technol. Inform. 192 (2013) 677–681.
[47] H. Xu, S.P. Stenner, S. Doan, K.B. Johnson, L.R. Waitman, J.C. Denny, MedEx: a medication information extraction system for clinical narratives, J. Am. Med. Inform. Assoc. 17 (2010) 19–24. doi:10.1197/jamia.M3378.
[48] S. Velupillai, D. Mowery, B.R. South, M. Kvist, H. Dalianis, Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis, Yearb. Med. Inform. 10 (2015) 183–193. doi:10.15265/IY-2015-009.
[49] A. Névéol, C. Grouin, J. Leixa, S. Rosset, P. Zweigenbaum, The Quaero French medical corpus: A ressource for medical entity recognition and normalization, in: Proc BioTextM Reyk., 2014.
[50] J.A. Kors, S. Clematide, S.A. Akhondi, E.M. van Mulligen, D. Rebholz-Schuhmann, A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC, J. Am. Med. Inform. Assoc. JAMIA. 22 (2015) 948–956. doi:10.1093/jamia/ocv037.
[51] L. Campillos, L. Deléger, C. Grouin, T. Hamon, A.-L. Ligozat, A. Névéol, A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT), Lang. Resour. Eval. (2017) 1–31. doi:10.1007/s10579-017-9382-y.
[52] A.-S. Jannot, E. Zapletal, P. Avillach, M.-F. Mamzer, A. Burgun, P. Degoulet, The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience, Int. J. Med. Inf. 102 (2017) 21–28. doi:10.1016/j.ijmedinf.2017.02.006.
[53] G. Hripcsak, D.J. Albers, Next-generation phenotyping of electronic health records, J. Am. Med. Inform. Assoc. JAMIA. 20 (2013) 117–121. doi:10.1136/amiajnl-2012-001145.
[54] J.C. Denny, Chapter 13: Mining Electronic Health Records in the Genomics Era, PLoS Comput. Biol. 8 (2012). doi:10.1371/journal.pcbi.1002823.
[55] V. Canuel, B. Rance, P. Avillach, P. Degoulet, A. Burgun, Translational research platforms integrating clinical and omics data: a review of publicly available solutions, Brief. Bioinform. 16 (2015) 280–290. doi:10.1093/bib/bbu006.
[56] D. Grande, N. Mitra, A. Shah, F. Wan, D.A. Asch, Public Preferences about Secondary Uses of Electronic Health Information, JAMA Intern. Med. 173 (2013) 1798–1806. doi:10.1001/jamainternmed.2013.9166.
[57] K. Spencer, C. Sanders, E.A. Whitley, D. Lund, J. Kaye, W.G. Dixon, Patient Perspectives on Sharing Anonymized Personal Health Data Using a Digital System for Dynamic Consent and Research Feedback: A Qualitative Study, J. Med. Internet Res. 18 (2016) e66. doi:10.2196/jmir.5011.

6. Discussion

Nous avons développé un entrepôt de données orienté document permettant d'explorer les données médicales d'un hôpital. Nous avons conçu le modèle de données et l'interface permettant d'interagir avec lui en considérant deux principales contraintes :

(1) Orienter cet entrepôt autour des comptes rendus médicaux, considérés comme l'unité la plus « humaine » et donc la plus proche des utilisateurs. En effet, les résultats biologiques sont eux mêmes souvent transmis sous forme de document synthétique mis en forme pour une lecture humaine plus aisée.

(2) Proposer des méthodes de data mining intégrées dans une interface web à l'usage des médecins dans leur quotidien. Cela signifie un travail important sur l'ergonomie de l'interface utilisateur ainsi que sur l'optimisation du temps de calcul.

Nous avons défini trois objectifs pour la réutilisation des données hospitalières :

(1) Rechercher des patients à partir du texte libre et des données structurées

(2) Explorer les données facilement pour réaliser du phénotypage haut débit

(3) Retrouver des patients par des méthodes de similarité pour le diagnostic ou le recrutement automatique

Le modèle de données

Ces deux contraintes et ces trois objectifs nous ont conduit à définir un schéma de base de données relationnel proche de celui d'i2b2. Notre particularité a été de séparer les données textuelles des données structurées. Dans notre modèle chaque donnée structurée est reliée à un compte rendu. Nous avons aussi volontairement séparé les données extraites du texte (concepts extraits) des données structurées de manière native dans les sources (PMSI, Biologie). Cette dichotomie est matérialisée dans la structure de la base de données dans une volonté de transparence quant à la validité des données vis à vis des utilisateurs biostatisticiens qui utiliseraient l'entrepôt directement.

La redondance des colonnes améliore la puissance de calcul en réduisant le nombre de jointures. Certains indicateurs sont calculés lors de la phase de chargement des données, comme par exemple, le nombre de patients par concept, l'âge du patient à la date des documents, le terme préféré utilisé dans les concepts extraits.

La recherche d'information

Notre approche de recherche en texte libre est innovante dans la mesure où la recherche n'est pas réduite à des concepts extraits du texte libre, mais bien à l'intégralité du texte libre. L'utilisateur peut rechercher des termes médicaux associés à des attributs et des qualitatifs

améliorant sa précision, ou ajouter des termes non médicaux dans sa recherche (diabete and instituteur or institutrice). Le logiciel Emerse (Hanauer et al., 2015) propose une approche similaire pour la recherche en texte libre mais il ne permet pas de filtrer automatiquement les faux positifs liés à la négation ou aux antécédents familiaux. Il ne permet pas non plus d'ajouter des critères structurés à la recherche.

Depuis 2011, i2b2 permet d'ajouter dans leur interface de requêtage structurée une recherche en texte libre (Murphy, 2011b). Le système de requêtage plein texte reste limité à un « like » SQL et ne permet pas l'expression d'une requête plus complexe rendu possible dans des technologies comme Lucene ("Apache Lucene - Apache Lucene Core," n.d.) ou Oracle Text ("Oracle Text," n.d.). Il est toutefois possible d'intégrer dans le modèle i2b2 la notion de contexte et de certitude en utilisant les *modifiers*. Néanmoins, en l'absence d'une interface dédiée pour vérifier les résultats obtenus par une recherche en texte libre, i2b2 n'est pas l'outil adapté à ce type de requête. Le challenge i2b2 en met à disposition des corpus annotés en anglais pour promouvoir l'extraction d'information à partir du texte libre : sur la détection du statut de fumeur du patient (Uzuner et al., 2008), sur la détection de l'obésité (Uzuner, 2009), sur la résolution des coréférences dans les textes (Uzuner et al., 2012), sur l'annotation de la temporalité (Sun et al., 2013b, 2013c), sur l'extraction des facteurs de risque de maladie cardiaque (Stubbs et al., 2015; Stubbs and Uzuner, 2015).

L'enrichissement terminologique réalisé sur les textes dans la base de données plutôt que sur la requête d'interrogation permet de conserver la complexité de cette requête (association de AND, OR, NOT, et éventuellement d'autres opérateurs spécifiques à Oracle Text : near, fuzzy). L'UMLS® Metathesaurus® nous permet d'enrichir les textes avec les synonymes des termes retrouvés. L'utilisation des relations « is a » nous a permis d'y ajouter les ancêtres de ces concepts et tous leurs synonymes. Nous avons réutilisé l'extraction des concepts dans les algorithmes de phénotypage et de similarité que nous avons décrites.

Vers une détection du contexte plus large

Notre capacité à séparer les phrases concernant les antécédents familiaux des phrases concernant le patient, améliore incontestablement la précision du moteur de recherche. La F-measure passe de 43% à 71% en ajoutant la détection des antécédents familiaux, puis à 91% en y ajoutant la détection de la négation. Nous pourrions enrichir les contextes détectés tels que le motif (d'hospitalisation, d'imagerie, de consultation), les antécédents du patient, la conclusion. La recherche d'information pourrait être ciblée sur un de ces contextes en particulier. De plus, les concepts extraits seraient alors plus finement classés. Ruch *et al.* présentent une méthode basée sur l'analyse du discours pour classer les phrases de résumés d'articles selon les quatre catégories objectif, méthode, résultat et conclusion(Ruch et al., 2007).

De la négation à la certitude : le doute, l'hypothèse, la recherche

Le niveau de certitude est traité sur deux niveaux « négatif » (certitude=-1) et « non négatif » (certitude=1). Nous devrons améliorer notre système afin de pouvoir ajouter les expressions de l'incertitude, de l'hypothèse, de la condition. Notre approche à base de règles d'expressions régulières nous a permis de générer un corpus de syntagmes considérés comme « négation » ou « non négation ». Il s'agit pour la suite de développer une interface de validation manuelle afin de créer un corpus validé permettant d'appliquer des méthodes de machine learning pour améliorer la détection des syntagmes négatifs (Wu et al., 2014), des syntagmes de doute et d'hypothèse.

Extraction phénotypique

Nous montrons la nécessité d'enrichir l'UMLS avec des terminologies françaises qui couvrent davantage le vocabulaire médical. Par exemple, seulement 1/4 des concepts HPO (terminologie utilisée pour la description phénotype des maladies rares dans Orphadata) ont un équivalent français dans l'UMLS (version 2017AA).

Nous allons intégrer dans les prochaines versions le type sémantique concernant l'anatomie humaine. Sans l'intégrer au phénotypage haut débit, cela permettra de réduire les concepts faux positifs liés à la présence du terme dans le nom d'un organe (« vésicule biliaire » et « vésicule »).

En utilisant les méthodes de Word Embedding, nous pensons pouvoir améliorer l'extraction de certains concepts qui sont souvent mal orthographiés dans les comptes rendus. Par exemple la recherche en texte libre de « Hirschsprung » renvoie 785 patients, « hirschprung » 354 patients, « hirchsprung » 18 patients et « hirshprung » 62 patients. En utilisant comme corpus d'apprentissage tous les comptes rendus contenant « Hirschsprung » nous pensons pouvoir indexer les comptes rendus contenant une faute d'orthographe dans le nom. Wu *et al.* a montré l'intérêt de cette méthode pour améliorer les méthodes d'extraction de termes dans les textes (Wu et al., 2015).

Vers un corpus annoté

Il devient indispensable de créer un corpus annoté en français pour d'une part développer des méthodes de machine learning, mais aussi plus généralement pour évaluer les algorithmes de traitement automatique du langage. A partir des extractions de concepts déjà réalisées, nous avons commencé à explorer la création de fichiers pré annotés compatible avec le logiciel d'annotation Brat développé par le MIT (Stenetorp et al., 2012). Il faudra prévoir une validation manuelle par un expert et éventuellement des ajouts.

Interopérabilité multi site

Le format XML choisi pour l'encapsulation des requêtes est un standard utilisé par la majorité des éditeurs. Toutefois le passage à une version JSON permettrait de réduire la volumétrie et de gagner en interopérabilité avec le langage Javascript (Nurseitov et al., 2009). L'interrogation multi sites par des entrepôts de données connectés est largement étudiée (SHRINE (Weber et al., 2009), ERH4CR (El Fadly et al., 2011), ConSoRe (Heudel et al., 2016)) et ouvre le data mining au delà du périmètre de l'établissement. Chaque initiative a développé son propre standard d'encapsulation des données (SHRINE Core Ontology (McMurry et al., 2013), EHR4CR Common Information Model (Daniel et al., 2016) basé sur FHIR, HL7 et les types sémantiques de l'UMLS). Il s'agira alors de modifier notre format d'encapsulation ou de développer un outil de transformation d'un standard à un autre pour rendre Dr Warehouse facilement interopérable avec les autres outils.

Le phénotypage haut débit:

Nous proposons un tri des concepts extraits à partir d'une cohorte de patients. Nous avons montré que nous obtenons une bonne précision sur les 50 premiers concepts sur les milliers de concepts extraits, et une excellente spécificité. Notre méthode est très simple, scalable et généralisable quelque soit la langue. Elle ne nécessite pas de corpus d'apprentissage ou de patients contrôles. Elle permet de réaliser une première exploration des données par les cliniciens à faible coût. Cependant, notre méthode est dépendante de l'algorithme d'extraction des concepts à partir du texte libre. L'amélioration de celle-ci aura un impact direct sur la description phénotypique automatisée.

Vers un PheWAS automatisé

Il serait judicieux d'étendre ce phénotypage aux données codées issues du codage CIM10 et des résultats de laboratoire. Des travaux ont été initiés dans ce sens par Antoine Neuraz afin de proposer une méthode d'appariement automatique de cas contrôle et d'intégrer ce PheWAS multimodal dans l'entrepôt de données.

Vers une histoire naturelle automatisée

Chaque concept extrait est associé à l'âge du patient à la date du compte rendu. Nous pensons pouvoir décrire l'histoire naturelle d'une maladie rare avec la moyenne d'âge de première apparition des phénotypes extraits. Par exemple, nous avons exploré la cohorte des patients atteints du Syndrome PI3kinase delta activée (Figure 3). L'évaluation est complexe car il n'existe pas de bases de connaissance formelles auxquelles se comparées.

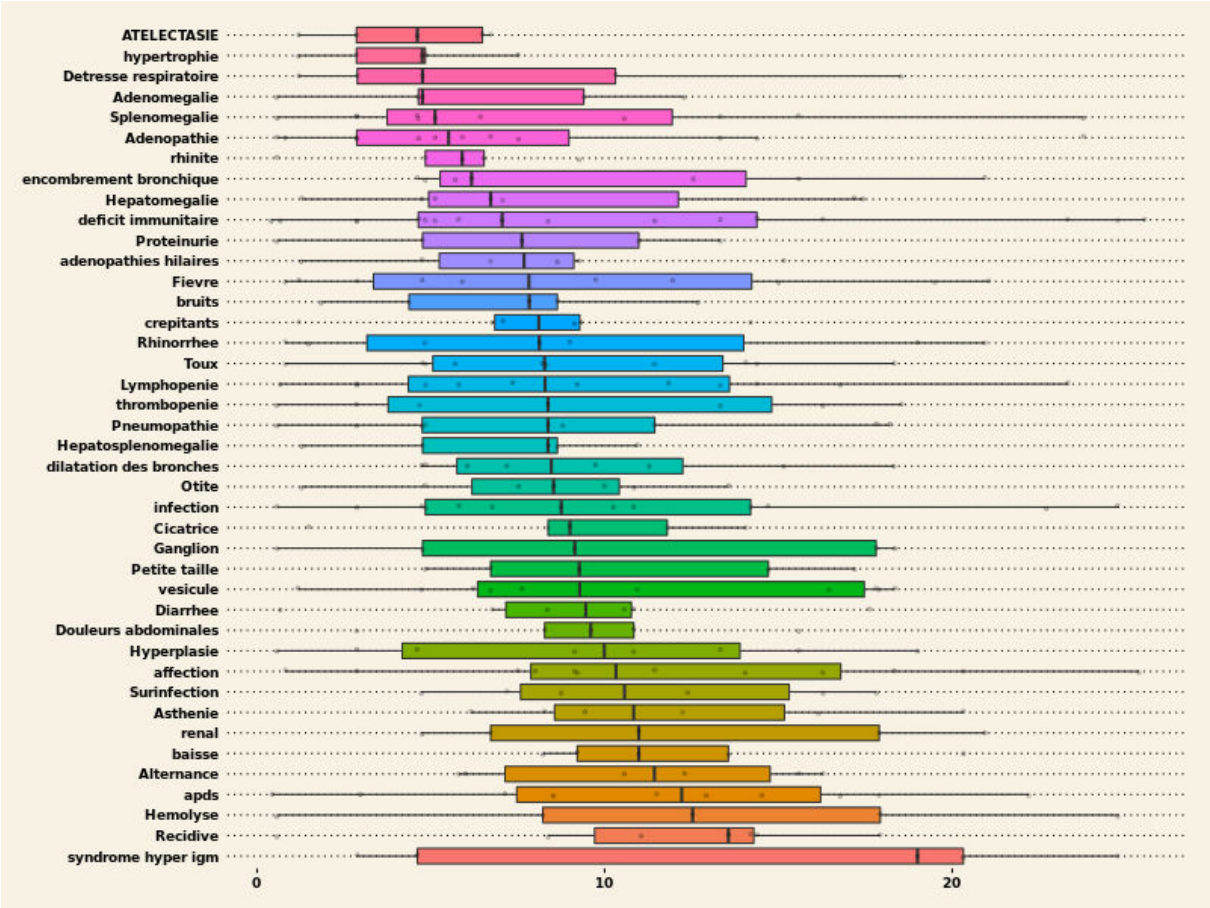


Figure 3 : Médiane de l'âge de première apparition des phénotypes associés au Syndrome PI3kinase delta activée

L'aide au diagnostic / l'aide au recrutement :

Nous avons présenté une méthode réutilisant l'extraction des concepts pour calculer des distances de similarité entre les patients. Nous avons utilisé le Vector Space Model pondéré par le TF-IDF.

Un travail important d'optimisation du temps de calcul a été nécessaire afin d'intégrer cette méthode dans une interface utilisateur. Pour cela, nous avons réduit l'espace vectoriel aux concepts du patient index. Et nous avons déterminé un nombre minimum de concepts (non négatifs) en commun afin de réduire le nombre de patients sur lesquels réaliser le calcul matriciel.

Notre évaluation montre une précision et un rappel en moyenne relativement bas (une précision à 0.51 et un rappel à 0.2) pour plusieurs raisons : le faible nombre de patients dans les cohortes, les faux positifs sont souvent considérés comme des diagnostics différentiels, les faux négatifs ont peu d'information clinique disponible. Les résultats obtenus restent toutefois pertinents dans le contexte des maladies rares.

Parmi les patients faux positifs, nous avons détecté un patient qui était retrouvé plusieurs fois comme similaire à des patients atteint du syndrome PI3K Delta Activé (APDS). Ce syndrome a été découvert en 2013 avec le gène *PIK3CD (Angulo et al., 2013)*, et enrichi en 2014 d'un

nouveau gène *PIK3R1 (Deau et al., 2014)*. La dernière visite de ce patient datait de 2012, il n'avait donc pas pu bénéficier d'un test génétique sur ces gènes. Nous l'avons signalé au médecin l'ayant pris en charge. Il s'est avéré par la suite que ce patient était effectivement muté sur le gène *PIK3R1*. Cet impact de Dr Warehouse sur le diagnostic et la prise en charge d'un patient, nous a convaincu de son intérêt dans l'hôpital, dans le quotidien des médecins.

Le calcul de similarité développé est perfectible notamment en réduisant le nombre de dimensions. Cette amélioration sera effectuée en intégrant la dépendance entre les concepts (relation entre signes et maladies : fièvre et grippe), et la distance sémantique entre les concepts (McInnes et al., 2009). Le calcul de similarité peut aussi prendre en paramètre d'autres données telles que l'âge du patient au moment des symptômes, les résultats biologiques (tendance, minimum, maximum).

Un autre axe de travail serait de calculer la similarité non plus à partir d'un patient index mais à partir d'une cohorte de patients index en prenant en compte les patients exclus par les experts lors de la revue manuelle des dossiers. L'ajout d'un algorithme de renforcement permettrait d'améliorer le rappel et la précision. Pasche *et al.* ont montré une amélioration de leur précision en modulant les concepts utilisés avec l'algorithme de Rocchio (Pasche et al., 2017; Salton, 1971).

Au delà de la détection de patients similaires permettant de rechercher des patients de l'entrepôt pouvant bénéficier d'un nouveau test génétique, nous pouvons utiliser cette méthode pour établir le diagnostic d'un patient index à partir des patients similaires. Nous avons intégré dans l'interface utilisateur la liste de tous les concepts des 20 patients similaires en éliminant ceux déjà présents pour le patient index. Un tri basé sur le TF-IDF permet d'afficher les concepts les plus pertinents parmi les 20 patients similaires. Une évaluation de ce système doit être réalisée, mais les résultats préliminaires sont prometteurs.

Nous avons développé une interface indépendante de Dr Warehouse, utilisant les sacs de termes (les sacs correspondants aux patients anonymisés). Un moteur de recherche permet de réaliser une recherche en texte libre sur ces sacs de termes. L'algorithme calcule les 50 concepts les plus représentatifs des patients retrouvés (TF-IDF) et affiche le résultat en précisant le type sémantique des concepts affichés. Notre approche data driven nous paraît particulièrement pertinente par le nombre de maladies rares prises en charge à l'hôpital Necker et ainsi décrites dans l'entrepôt de données. Le système, une fois évalué, pourra être utilisé comme système d'aide au diagnostic à Necker et dans les autres hôpitaux. Il pourrait permettre de réduire le délai avant le diagnostic pour les patients atteints de maladies rares, délai estimé à 7.6 années avec une moyenne de 8 médecins consultés selon l'étude Shire ("The Shire Rare Disease Impact Report (2013 – US and UK population)," n.d.).

Une ergonomie adaptée

L'ergonomie des interface homme-machine a largement été décrite dans la littérature (Patel and Kushniruk, 1998; Shneiderman, 1997). Nous avons intégré les algorithmes dans le logiciel Dr Warehouse en centrant le développement autour des utilisateurs (médecins et attachés de recherche clinique) (McCracken et al., 2003), en adaptant les interfaces par des évaluations itératives (Nielsen, 1993; Patel and Kushniruk, 1998) et par un développement basé sur la méthode Agile (Cockburn, 2002). Le projet VISAGE a utilisé l'indicateur « nombre de clics » comme critère d'évaluation pour comparer leur ergonomie avec celle d'i2b2 (Zhang et al., 2010). Ils ont évalué qu'il fallait 2 à 4 fois plus de clics pour réaliser une recherche dans i2b2 que dans leur application et que ça prenait 3 à 7 fois plus de temps.

Pour adapter l'interface à la recherche en texte libre, Dr Warehouse s'inspire de Google (Brin and Page, 1998) et affiche un extrait du document contenant les critères recherchés (aussi bien le texte libre, ses synonymes et ses hyponymes que les requêtes structurées). Indispensable pour valider les patients, cet aperçu du compte rendu participe à l'adhésion des utilisateurs à la recherche textuelle malgré la possibilité de retrouver des patients faux positifs.

Le rappel et la précision sont les deux principales mesures de performance à optimiser pour un moteur de recherche (Rijsbergen, 1979). Dans le cas d'un entrepôt de données biomédicales, plus particulièrement dans le contexte des maladies rares, il semble avantageux de pouvoir diminuer la précision ou diminuer le rappel en fonction des résultats et de l'objectif. Dans le cadre d'une association phénotypique extrêmement rare, on souhaitera vérifier tous les patients qui ont l'ensemble des signes évoqués dans leurs comptes rendus, même s'il s'agit d'une suspicion, voire d'une négation, ou d'un antécédent familial. Le modèle de données et l'interface tels qu'ils sont construits permettent de rechercher dans l'ensemble des comptes rendus, voire même dans les négations spécifiquement, ou dans les antécédents familiaux. Les moteurs de recherche intégrant la détection de la négation pour la recherche en texte libre ne permettent pas d'étendre la recherche aux données considérées comme absentes. Par exemple, Martinez *et al.* utilisent NegEx et excluent totalement les termes niés pour la recherche d'information (Martinez et al., 2014). Les entrepôts de données basées sur la technologie OLAP ne permettent pas non plus d'utiliser ces données car ils s'orientent davantage vers des calculs d'indicateurs.

Le dossier patient informatisé, une source de données à optimiser :

Hripcsak (Hripcsak and Albers, 2013) évoque la nécessité de bien appréhender le contenu du DPI pour développer des outils pertinents. Il insiste sur le besoin d'améliorer la production de soin à la source (coder, biais, données manquantes etc.). Une étude de 2016 montre qu'en moyenne un médecin passe 49% de son temps à saisir des données dans le dossier patient (Sinsky et al., 2016). Malgré l'importance de cette tâche, le médecin doit pouvoir passer

davantage de temps en face à face avec le patient. Des initiatives de scribes ont montré gain obtenu non négligeable aussi bien quant à la tenue du dossier, qu'au temps gagné et qu'au nombre de patients vus qui a augmenté (Bank et al., 2013). La solution la plus efficace serait de proposer des dossiers patients informatisés intelligents, intégrant un codage semi automatique, l'ajout automatique de bloc pertinent en fonction de l'histoire du patient. Mais, le temps nécessaire à la documentation du patient sera, à un certain stade, incompressible, car nécessaire au processus de diagnostic du médecin.

Les règles éthiques

De part la nature des données manipulées et la puissance des outils de fouille de données, il est indispensable de disposer d'un encadrement éthique de cet entrepôt de données. Le développement de Dr Warehouse a été accompagné par un conseil scientifique dont un groupe de travail était dédié aux aspects éthiques et juridiques.

Une charte d'utilisation a été définie, et présentée aux instances éthiques de l'hôpital Necker-Enfants Malades (Conseil d'Ethique de Necker Enfants Malades, Association des usagers de l'hôpital) (Lamas et al., 2015). Une réflexion collective s'est engagée au sein de Necker et de l'Institut Imagine pour créer un environnement adapté au soin et à la recherche sur les maladies rares. Ceci a permis une contribution au débat national conduit par la Commission nationale de l'informatique et des libertés (CNIL) (Garcelon, 2017).

Le paramétrage de Dr Warehouse permet de prendre en compte la plupart des règles d'accès de l'établissement dans lequel il pourrait être installé. La personne en charge des droits d'accès peut définir des profils avec droits spécifiques pour chaque fonctionnalité, d'accès aux données nominatives ou anonymisées, aux données par périmètre de soin (service, hôpital, aucun), aux données par source de données.

Aujourd'hui, Dr Warehouse n'est ouvert qu'aux médecins prenant en charge les patients.

Pour une ouverture de l'accès aux données aux chercheurs, il faudrait mettre en place le plus tôt possible un système de consentement électronique afin de faciliter les études rétrospectives qui pourraient nécessité l'utilisation de matériel génétique. Grande *et al.* ont montré le fort potentiel de cette méthode pour une meilleure acceptation par les patients (Grande et al., 2013).

Nous réfléchissons par ailleurs à l'opportunité de développer une plateforme d'accès à l'entrepôt de données. La Vanderbilt University a développé la plateforme STARBRITE afin de gérer les études et les accès à leur entrepôt et plus généralement à toutes les ressources disponibles pour la recherche (assistance, éthique, guide, expertise etc.) (Harris et al., 2011). L'Institut Imagine et Necker pourraient tirer bénéfice d'une plateforme similaire pour faciliter les nombreux projets de recherche conduits sur le site.

Dissémination

Dr Warehouse est sous licence open source GPL GNU v3. Il est en téléchargement libre sur git https://github.com/imagine-bdd/DRWH .

Une démonstration est disponible en accès libre https://imagine-plateforme-bdd.fr/dwh_pubmed. Les 4 639 patients fictifs ont été générés à partir du site *Fake Name Generator*, les 100 867 comptes rendus sont des résumés de Pubmed concernant des maladies rares. Les résultats biologiques sont des résultats semi aléatoires de données de Necker (afin de conserver une cohérence sur la valeur d'un examen), les parcours de soin sont aussi des parcours semi aléatoires générés à partir des données de Necker. La détection de la négation et des antécédents familiaux n'a pas été appliquée sur cette version de démonstration car notre algorithme ne fonctionne pour l'instant que pour le français.

7. Conclusion

Dans le contexte des maladies rares, la standardisation du dossier patient informatisé est éloignée de la pratique médicale qui nécessite une description exhaustive du tableau clinique du patient et de ses antécédents familiaux. La réutilisation du texte libre pour la recherche ne peut pas se faire avec les mêmes outils que ceux utilisés pour les données structurées à la source.

Le développement de méthodes et d'interfaces adaptées aux données textuelles est un moyen de remettre la recherche translationnelle au lit du malade. Le clinicien se réapproprie ses données cliniques dans ce nouveau paradigme de fouille de données. Ses données deviennent une nouvelle source de connaissance pouvant l'aider dans la prise en charge de ses patients.

Nous avons développé Dr Warehouse dans cet objectif de remettre le document clinique au cœur du dispositif de la recherche translationnelle. Notre travail relève les challenges identifiés par Prokosch en proposant un entrepôt de données clinique complet, des méthodes de phénotypage automatique et d'aide au diagnostic.

Ces méthodes pourront être encore améliorées par l'apport du machine learning. Mais cela nécessite d'avoir des corpus de comptes rendus cliniques annotés en français comme l'a souligné Meystre (Meystre et al., 2008).

Dr Warehouse a été déployé récemment à l'hôpital Necker, nous constatons déjà une adhésion quasiment unanime à ce logiciel par les cliniciens et les assistants de recherche clinique. Son utilisation reste aujourd'hui essentiellement basée sur la recherche de patients et l'exploration du dossier d'un patient. Des séances de formation et l'implication du département d'informatique médicale devraient progressivement faire entrer dans les habitudes des cliniciens ces nouveaux usages de leurs données cliniques.

Bibliographie

Al-Awqati, Q., 2006. How to write a case report: lessons from 1600 B.C. Kidney Int. 69, 2113–2114. doi:10.1038/sj.ki.5001592

Alonso, O., Gertz, M., Baeza-Yates, R., 2007. On the Value of Temporal Information in Information Retrieval. SIGIR Forum 41, 35–41. doi:10.1145/1328964.1328968

Angulo, I., Vadas, O., Garçon, F., Banham-Hall, E., Plagnol, V., Leahy, T.R., Baxendale, H., Coulter, T., Curtis, J., Wu, C., Blake-Palmer, K., Perisic, O., Smyth, D., Maes, M., Fiddler, C., Juss, J., Cilliers, D., Markelj, G., Chandra, A., Farmer, G., Kielkowska, A., Clark, J., Kracker, S., Debré, M., Picard, C., Pellier, I., Jabado, N., Morris, J.A., Barcenas-Morales, G., Fischer, A., Stephens, L., Hawkins, P., Barrett, J.C., Abinun, M., Clatworthy, M., Durandy, A., Doffinger, R., Chilvers, E.R., Cant, A.J., Kumararatne, D., Okkenhaug, K., Williams, R.L., Condliffe, A., Nejentsev, S., 2013. Phosphoinositide 3-kinase δ gene mutation predisposes to respiratory infection and airway damage. Science 342, 866–871. doi:10.1126/science.1243292

Apache Lucene - Apache Lucene Core [WWW Document], n.d. URL https://lucene.apache.org/core/ (accessed 9.10.17).

Aronson, A.R., 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc. AMIA Symp. 17.

Bank, A.J., Obetz, C., Konrardy, A., Khan, A., Pillai, K.M., McKinley, B.J., Gage, R.M., Turnbull, M.A., Kenney, W.O., 2013. Impact of scribes on patient interaction, productivity, and revenue in a cardiology clinic: a prospective study. Clin. Outcomes Res. CEOR 5, 399–406. doi:10.2147/CEOR.S49010

Barrows Jr, R.C., Busuioc, M., Friedman, C., 2000. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. Proc. AMIA Symp. 51–55.

Bashyam, V., Divita, G., Bennett, D.B., Browne, A.C., Taira, R.K., 2007. A normalized lexical lookup approach to identifying UMLS concepts in free text. Stud. Health Technol. Inform. 129, 545–549.

Berndt, D.J., Hevner, A.R., Studnicki, J., 1998. CATCH/IT: a data warehouse to support comprehensive assessment for tracking community health. Proc. AMIA Symp. 250–254.

Brammen, D., Katzer, C., Röhrig, R., Weismüller, K., Maier, M., Hossain, H., Menges, T., Hempelmann, G., Chakraborty, T., 2005. An integrated data-warehouse-concept for clinical and biological information. Stud. Health Technol. Inform. 116, 9–14.

Brandt, C.A., Morse, R., Matthews, K., Sun, K., Deshpande, A.M., Gadagkar, R., Cohen, D.B., Miller, P.L., Nadkarni, P.M., 2002. Metadata-driven creation of data marts from an EAV-modeled clinical research database. Int. J. Med. Inf. 65, 225–241.

Brin, S., Page, L., 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine, in: Proceedings of the Seventh International Conference on World Wide Web 7, WWW7. Elsevier

Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, pp. 107–117.

Brooks, C.J., Stephens, J.W., Price, D.E., Ford, D.V., Lyons, R.A., Prior, S.L., Bain, S.C., 2009. Use of a patient linked data warehouse to facilitate diabetes trial recruitment from primary care. Prim. Care Diabetes 3, 245–248. doi:10.1016/j.pcd.2009.06.004

Buske, O.J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., Misyura, A., Friedman, T., Beaulieu, C., Bone, W.P., Links, A.E., Washington, N.L., Haendel, M.A., Robinson, P.N., Boerkoel, C.F., Adams, D., Gahl, W.A., Boycott, K.M., Brudno, M., 2015. PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. Hum. Mutat. 36, 931–940. doi:10.1002/humu.22851

Campbell, D.A., Johnson, S.B., 2002. A Transformational-based Learner for Dependency Grammars in Discharge Summaries, in: Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain - Volume 3, BioMed '02. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 37–44. doi:10.3115/1118149.1118155

Campbell, D.A., Johnson, S.B., 2001. Comparing syntactic complexity in medical and non-medical corpora. Proc. AMIA Symp. 90–94.

Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G., 2001a. A simple algorithm for identifying negated findings and diseases in discharge summaries. J. Biomed. Inform. 34, 301–310. doi:10.1006/jbin.2001.1029

Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G., 2001b. Evaluation of negation phrases in narrative clinical reports. Proc. AMIA Symp. 2001, 105–109.

Chapman, W.W., Hillert, D., Velupillai, S., Kvist, M., Skeppstedt, M., Chapman, B.E., Conway, M., Tharp, M., Mowery, D.L., Deleger, L., 2013. Extending the NegEx lexicon for multiple languages. Stud. Health Technol. Inform. 192, 677–681.

Charon, R., 2012. At the membranes of care: stories in narrative medicine. Acad. Med. J. Assoc. Am. Med. Coll. 87, 342–347. doi:10.1097/ACM.0b013e3182446fbb

Chen, C.-K., Mungall, C.J., Gkoutos, G.V., Doelken, S.C., Köhler, S., Ruef, B.J., Smith, C., Westerfield, M., Robinson, P.N., Lewis, S.E., Schofield, P.N., Smedley, D., 2012. MouseFinder: candidate disease genes from mouse phenotype data. Hum. Mutat. 33, 858–866. doi:10.1002/humu.22051

Chen, Z., 2001. Intelligent Data Warehousing: From Data Preparation to Data Mining. CRC Press, Inc., Boca Raton, FL, USA.

Chu, D., Dowling, J.N., Chapman, W.W., 2006. Evaluating the effectiveness of four contextual features in classifying annotated clinical conditions in emergency department reports. AMIA Annu. Symp. Proc. AMIA Symp. 141–145.

Chute, C.G., Beck, S.A., Fisk, T.B., Mohr, D.N., 2010. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J. Am. Med. Inform. Assoc. 17, 131–135. doi:10.1136/jamia.2009.002691

Clark, C., Aberdeen, J., Coarr, M., Tresner-Kirsch, D., Wellner, B., Yeh, A., Hirschman, L., 2011.

MITRE system for clinical assertion status classification. J. Am. Med. Inform. Assoc. JAMIA 18, 563. doi:10.1136/amiajnl-2011-000164

Cockburn, A., 2002. Agile Software Development. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Codd, E.F., 1970. A Relational Model of Data for Large Shared Data Banks. Commun ACM 13, 377–387. doi:10.1145/362384.362685

Code de la santé publique - Article R1112-2, n.d. , Code de la santé publique.

Connolly, T.M., Begg, C.E., 2004. Database Systems: A Practical Approach to Design, Implementation and Management (4th Edition). Pearson Addison Wesley.

Crowley, R.S., Castine, M., Mitchell, K., Chavan, G., McSherry, T., Feldman, M., 2010. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. J. Am. Med. Inform. Assoc. JAMIA 17, 253–264. doi:10.1136/jamia.2009.002295

Cuggia, M., Bayat, S., Garcelon, N., Sanders, L., Rouget, F., Coursin, A., Pladys, P., 2010. A full-text information retrieval system for an epidemiological registry. Stud. Health Technol. Inform. 160, 491–495.

Danciu, I., Cowan, J.D., Basford, M., Wang, X., Saip, A., Osgood, S., Shirey-Rice, J., Kirby, J., Harris, P.A., 2014. Secondary use of clinical data: The Vanderbilt approach. J. Biomed. Inform., Special Section: Methods in Clinical Research Informatics 52, 28–35. doi:10.1016/j.jbi.2014.02.003

Daniel, C., Ouagne, D., Sadou, E., Forsberg, K., Gilchrist, M.M., Zapletal, E., Paris, N., Hussain, S., Jaulent, M.-C., MD, D.K., 2016. Cross border semantic interoperability for clinical research: the EHR4CR semantic resources and services. AMIA Summits Transl. Sci. Proc. 2016, 51–59.

Deau, M.-C., Heurtier, L., Frange, P., Suarez, F., Bole-Feysot, C., Nitschke, P., Cavazzana, M., Picard, C., Durandy, A., Fischer, A., Kracker, S., 2014. A human immunodeficiency caused by mutations in the PIK3R1 gene. J. Clin. Invest. 124, 3923–3928. doi:10.1172/JCI75746

Deshmukh, V.G., Meystre, S.M., Mitchell, J.A., 2009. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC Med. Res. Methodol. 9, 70. doi:10.1186/1471-2288-9-70

DGOS, 2017. Atlas des systèmes d'information hospitaliers [WWW Document]. Ministère Solidar. Santé. URL http://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/e-sante/sih/article/atlas-des-systemes-d-information-hospitaliers (accessed 9.5.17).

Ding, J., Erdal, S., Dhaval, R., Kamal, J., 2007. Augmenting Oracle Text with the UMLS for enhanced searching of free-text medical reports. AMIA Annu. Symp. Proc. AMIA Symp. 940.

Dinu, V., Nadkarni, P., 2007. Guidelines for the Effective Use of Entity-Attribute-Value Modeling for Biomedical Databases. Int. J. Med. Inf. 76, 769–779. doi:10.1016/j.ijmedinf.2006.09.023

Divita, G., Zeng, Q.T., Gundlapalli, A.V., Duvall, S., Nebeker, J., Samore, M.H., 2014. Sophia: A

Expedient UMLS Concept Extraction Annotator. AMIA. Annu. Symp. Proc. 2014, 467–476.

Dusetzina, S.B., Tyree, S., Meyer, A.-M., Meyer, A., Green, L., Carpenter, W.R., 2014. An Overview of Record Linkage Methods. Agency for Healthcare Research and Quality (US).

El Fadly, A., Rance, B., Lucas, N., Mead, C., Chatellier, G., Lastic, P.-Y., Jaulent, M.-C., Daniel, C., 2011. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. J. Biomed. Inform. 44 Suppl 1, S94-102. doi:10.1016/j.jbi.2011.07.007

Erinjeri, J.P., Picus, D., Prior, F.W., Rubin, D.A., Koppel, P., 2009. Development of a Google-Based Search Engine for Data Mining Radiology Reports. J. Digit. Imaging Off. J. Soc. Comput. Appl. Radiol. 22, 348–356. doi:10.1007/s10278-008-9110-7

Escudié, J.-B., Jannot, A.-S., Zapletal, E., Cohen, S., Malamut, G., Burgun, A., Rance, B., 2015. Reviewing 741 patients records in two hours with FASTVISU. AMIA. Annu. Symp. Proc. 2015, 553–559.

Fact SheetSPECIALIST Lexicon [WWW Document], n.d. URL https://www.nlm.nih.gov/pubs/factsheets/umlslex.html (accessed 9.23.17).

Fact SheetUMLS® Metathesaurus® [WWW Document], n.d. URL https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html (accessed 9.23.17).

Fact SheetUMLS® Semantic Network [WWW Document], n.d. URL https://www.nlm.nih.gov/pubs/factsheets/umlssemn.html (accessed 9.23.17).

Finkel, J.R., Grenager, T., Manning, C., 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 363–370. doi:10.3115/1219840.1219885

Frankovich, J., Longhurst, C.A., Sutherland, S.M., 2011. Evidence-based medicine in the EMR era. N. Engl. J. Med. 365, 1758–1759. doi:10.1056/NEJMp1108726

Friedlin, J., McDonald, C.J., 2006. Using A Natural Language Processing System to Extract and Code Family History Data from Admission Reports. AMIA. Annu. Symp. Proc. 2006, 925.

Friedman, C., 1997. Towards a comprehensive medical language processing system: methods and issues. Proc. AMIA Annu. Fall Symp. 595–599.

Friedman, C., Shagina, L., Lussier, Y., Hripcsak, G., 2004. Automated encoding of clinical documents based on natural language processing. J. Am. Med. Inform. Assoc. JAMIA 11, 392–402. doi:10.1197/jamia.M1552

Garcelon, N., 2017. Journée sur la recherche en santé dans ses aspects éthiques et réglementaires (données, algorithmes). Débat public initié par la CNIL sur les enjeux éthiques et de société soulevés par les algorithmes et l'Intelligence Artificielle (IA).

Garcelon, N., Courteille, V., Fischer, A., Mahlaoui, N., 2014. Epidemiology of PID: Innovative New Way to Identify Patients in the CEREDIH Registry Through a Medical Data Warehouse. J. Clin.

Immunol. 34, S361–S362.

Gillum, R.F., 2013. From papyrus to the electronic tablet: a brief history of the clinical medical record with lessons for the digital age. Am. J. Med. 126, 853–857. doi:10.1016/j.amjmed.2013.03.024

Goryachev, S., Sordo, M., Zeng, Q., Ngo, L., 2006. Implementation and evaluation of four different methods of negation detection. Boston MA DSG.

Gottlieb, M.M., Arenillas, D.J., Maithripala, S., Maurer, Z.D., Tarailo Graovac, M., Armstrong, L., Patel, M., van Karnebeek, C., Wasserman, W.W., 2015. GeneYenta: a phenotype-based rare disease case matching tool based on online dating algorithms for the acceleration of exome interpretation. Hum. Mutat. 36, 432–438. doi:10.1002/humu.22772

Grabar, N., Varoutas, P.-C., Rizand, P., Livartowski, A., Hamon, T., 2008. Automatic acquisition of synonyms from French UMLS for enhanced search of EHRs. Stud. Health Technol. Inform. 136, 809–814.

Grande, D., Mitra, N., Shah, A., Wan, F., Asch, D.A., 2013. Public Preferences about Secondary Uses of Electronic Health Information. JAMA Intern. Med. 173, 1798–1806. doi:10.1001/jamainternmed.2013.9166

Grannis, S.J., Overhage, J.M., McDonald, C.J., 2002. Analysis of identifier performance using a deterministic linkage algorithm. Proc. AMIA Symp. 305–309.

Griffon, N., Chebil, W., Rollin, L., Kerdelhue, G., Thirion, B., Gehanno, J.-F., Darmoni, S.J., 2012. Performance evaluation of Unified Medical Language System®'s synonyms expansion to query PubMed. BMC Med. Inform. Decis. Mak. 12, 12. doi:10.1186/1472-6947-12-12

Groza, T., Köhler, S., Moldenhauer, D., Vasilevsky, N., Baynam, G., Zemojtel, T., Schriml, L.M., Kibbe, W.A., Schofield, P.N., Beck, T., Vasant, D., Brookes, A.J., Zankl, A., Washington, N.L., Mungall, C.J., Lewis, S.E., Haendel, M.A., Parkinson, H., Robinson, P.N., 2015. The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. Am. J. Hum. Genet. 97, 111–124. doi:10.1016/j.ajhg.2015.05.020

Groza, T., Verspoor, K., 2015. Assessing the Impact of Case Sensitivity and Term Information Gain on Biomedical Concept Recognition. PLoS ONE 10. doi:10.1371/journal.pone.0119091

Hanauer, D.A., 2006. EMERSE: The Electronic Medical Record Search Engine. AMIA. Annu. Symp. Proc. 2006, 941.

Hanauer, D.A., Mei, Q., Law, J., Khanna, R., Zheng, K., 2015. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). J. Biomed. Inform. 55, 290–300. doi:10.1016/j.jbi.2015.05.003

Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W., 2009. Context: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. J. Biomed. Inform. 42, 839–851. doi:10.1016/j.jbi.2009.05.002

Harris, P.A., Swafford, J.A., Edwards, T.L., Zhang, M., Nigavekar, S.S., Yarbrough, T.R., Lane, L.D.,

Helmer, T., Lebo, L.A., Mayo, G., Masys, D.R., Bernard, G.R., Pulley, J.M., 2011. StarBRITE: The Vanderbilt University Biomedical Research Integration, Translation and Education Portal. J. Biomed. Inform. 44, 655–662. doi:10.1016/j.jbi.2011.01.014

Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J.G., 2009. Research Electronic Data Capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. 42, 377–381. doi:10.1016/j.jbi.2008.08.010

Heinze, J., 2014. History of Business Intelligence [WWW Document]. Better Buys. URL https://www.betterbuys.com/bi/history-of-business-intelligence/ (accessed 9.2.17).

Hess, V., 2010. [Formalizing observation: The emergence of the modern patient record exemplified by Berlin and Paris medicine, 1725-1830]. Medizinhist. J. 45, 293–340.

Heudel, P., Livartowski, A., Arveux, P., Willm, E., Jamain, C., 2016. [The ConSoRe project supports the implementation of big data in oncology]. Bull. Cancer (Paris) 103, 949–950. doi:10.1016/j.bulcan.2016.10.001

Home - Gene - NCBI [WWW Document], n.d. URL https://www.ncbi.nlm.nih.gov/gene (accessed 9.25.17).

Hripcsak, G., Albers, D.J., 2013. Next-generation phenotyping of electronic health records. J. Am. Med. Inform. Assoc. JAMIA 20, 117–121. doi:10.1136/amiajnl-2012-001145

Hu, H., Correll, M., Kvecher, L., Osmond, M., Clark, J., Bekhash, A., Schwab, G., Gao, D., Gao, J., Kubatin, V., Shriver, C.D., Hooke, J.A., Maxwell, L.G., Kovatich, A.J., Sheldon, J.G., Liebman, M.N., Mural, R.J., 2011. DW4TR: A Data Warehouse for Translational Research. J. Biomed. Inform. 44, 1004–1019. doi:10.1016/j.jbi.2011.08.003

Huang, Y., Lowe, H.J., 2007. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. J. Am. Med. Inform. Assoc. JAMIA 14, 304–311. doi:10.1197/jamia.M2284

Inmon, B., 2009. What a Data Warehouse is Not by Bill Inmon - BeyeNETWORK [WWW Document]. BeyeNETWORK. URL http://www.b-eye-network.com/view/11352 (accessed 8.20.17).

Inmon, W.H., 1992. Building the Data Warehouse. John Wiley & Sons, Inc., New York, NY, USA.

INSERM, 1997. Orphadata: Free access data from Orphanet. © INSERM 1997. Available on http://www.orphadata.org. Data version (XML data version) [WWW Document]. URL http://www.orphadata.org/cgi-bin/inc/product4.inc.php (accessed 9.24.17).

Introducing the Necker-Enfants Malades hospital, 2015. . Hôp. Necker-Enfants Mal.

Jannot, A.-S., Zapletal, E., Avillach, P., Mamzer, M.-F., Burgun, A., Degoulet, P., 2017. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. Int. J. Med. Inf. 102, 21–28. doi:10.1016/j.ijmedinf.2017.02.006

Jones, K.S., 1972. A statistical interpretation of term specificity and its application in retrieval. J. Doc. 28, 11–21.

Kalra, D., Beale, T., Heard, S., 2005. The openEHR Foundation. Stud. Health Technol. Inform. 115, 153–173.

Kho, A.N., Pacheco, J.A., Peissig, P.L., Rasmussen, L., Newton, K.M., Weston, N., Crane, P.K., Pathak, J., Chute, C.G., Bielinski, S.J., Kullo, I.J., Li, R., Manolio, T.A., Chisholm, R.L., Denny, J.C., 2011. Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. Sci. Transl. Med. 3, 79re1. doi:10.1126/scitranslmed.3001807

King, B., Wang, L., Provalov, I., Zhou, J., 2011. The twentieth Text REtrieval Conference Proceedings (TREC) 2011. Presented at the Cengage Learning at TREC 2011 medical track.

Klann, J.G., Abend, A., Raghavan, V.A., Mandl, K.D., Murphy, S.N., 2016. Data interchange using i2b2. J. Am. Med. Inform. Assoc. JAMIA 23, 909–915. doi:10.1093/jamia/ocv188

Köhncke, B., Siehndel, P., Balke, W.-T., 2013. Bridging the Gap — Using External Knowledge Bases for Context-Aware Document Retrieval, in: Proceedings of the 15th International Conference on Digital Libraries: Social Media and Community Networks - Volume 8279, ICADL 2013. Springer-Verlag New York, Inc., New York, NY, USA, pp. 11–20. doi:10.1007/978-3-319-03599-4_2

Kortüm, K.U., Müller, M., Kern, C., Babenko, A., Mayer, W.J., Kampik, A., Kreutzer, T.C., Priglinger, S., Hirneiss, C., 2017. Using electronic health records to build an ophthalmological data warehouse and visualize patients' data. Am. J. Ophthalmol. doi:10.1016/j.ajo.2017.03.026

Krasowski, M.D., Schriever, A., Mathur, G., Blau, J.L., Stauffer, S.L., Ford, B.A., 2015. Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research. J. Pathol. Inform. 6, 45. doi:10.4103/2153-3539.161615

Lamas, E., Barh, A., Brown, D., Jaulent, M.-C., 2015. Ethical, Legal and Social Issues related to the health data-warehouses: re-using health data in the research and public health research. Stud. Health Technol. Inform. 210, 719–723.

Lee, J., Maslove, D.M., Dubin, J.A., 2015. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. PloS One 10, e0127428. doi:10.1371/journal.pone.0127428

Lewis, N., Gruhl, D., Yang, H., 2011. Extracting Family History Diagnosis from Clinical Texts. Presented at the Conference on Bioinformatics and Computational Biology, BICoB-2011, New Orleans, Louisiana, USA.

Lindberg, D.A., Humphreys, B.L., McCray, A.T., 1993. The Unified Medical Language System. Methods Inf. Med. 32, 281–291.

L'Institut Imagine [WWW Document], n.d. URL http://www.institutimagine.org/fr/l-institut-imagine.html (accessed 9.23.17).

Livartowski, A., 2016. Un Google 3.0 du cancer : est-ce possible ? — Challenge4Cancer [WWW Document]. URL http://wiki.epidemium.cc/wiki/Un_Google_3.0_du_cancer_:_est-ce_possible_%3F (accessed 9.23.17).

Long, W., 2005. Extracting Diagnoses from Discharge Summaries. AMIA. Annu. Symp. Proc. 2005, 470–474.

Lowe, H.J., Ferris, T.A., Hernandez, P.M., Weber, S.C., 2009. STRIDE--An integrated standards-based translational research informatics platform. AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp. 2009, 391–395.

Marco-Ruiz, L., Moner, D., Maldonado, J.A., Kolstrup, N., Bellika, J.G., 2015. Archetype-based data warehouse environment to enable the reuse of electronic health record data. Int. J. Med. Inf. 84, 702–714. doi:10.1016/j.ijmedinf.2015.05.016

Martinez, D., Otegi, A., Soroa, A., Agirre, E., 2014. Improving search over Electronic Health Records using UMLS-based query expansion through random walks. J. Biomed. Inform. 51, 100–106. doi:10.1016/j.jbi.2014.04.013

McCowan, C., Thomson, E., Szmigielski, C.A., Kalra, D., Sullivan, F.M., Prokosch, H.-U., Dugas, M., Ford, I., 2015. Using Electronic Health Records to Support Clinical Trials: A Report on Stakeholder Engagement for EHR4CR. BioMed Res. Int. 2015, 707891. doi:10.1155/2015/707891

McCracken, D.D., Spool, J.M., Wolfe, R.J., 2003. User-Centered Web Site Development: A Human-Computer Interaction Approach. Pearson Education.

McCray, A.T., Burgun, A., Bodenreider, O., 2001. Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. Stud. Health Technol. Inform. 84, 216–220.

McInnes, B.T., Pedersen, T., Pakhomov, S.V.S., 2009. UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. AMIA. Annu. Symp. Proc. 2009, 431–435.

McMurry, A.J., Murphy, S.N., MacFadden, D., Weber, G., Simons, W.W., Orechia, J., Bickel, J., Wattanasin, N., Gilbert, C., Trevvett, P., Churchill, S., Kohane, I.S., 2013. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. PLoS ONE 8. doi:10.1371/journal.pone.0055811

Mehrabi, S., Krishnan, A., Sohn, S., Roch, A.M., Schmidt, H., Kesterson, J., Beesley, C., Dexter, P., Max Schmidt, C., Liu, H., Palakal, M., 2015. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. J. Biomed. Inform. 54, 213–219. doi:10.1016/j.jbi.2015.02.010

Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., 2008. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb. Med. Inform. 128–144.

Montani, S., Portinale, L., Leonardi, G., Bellazzi, R., Bellazzi, R., 2006. Case-based Retrieval to Support the Treatment of End Stage Renal Failure Patients. Artif Intell Med 37, 31–42. doi:10.1016/j.artmed.2005.06.003

Müller, H., Clough, P., Deselaers, T., Caputo, B., 2010. ImageCLEF: Experimental Evaluation in Visual Information Retrieval, 1st ed. Springer Publishing Company, Incorporated.

Murphy, S., 2011a. Modifiers in i2b2 Data Model - i2b2 Developer's Forum - i2b2 Community Wiki [WWW Document]. URL https://community.i2b2.org/wiki/display/DevForum/Modifiers+in+i2b2+Data+Model (accessed

8.20.17).

Murphy, S., 2011b. Text search in i2b2 - i2b2 Developer's Forum - i2b2 Community Wiki [WWW Document]. URL https://community.i2b2.org/wiki/display/DevForum/Text+search+in+i2b2 (accessed 9.21.17).

Murphy, S., Barnett, G., Chueh, H., 2000. Visual query tool for finding patient cohorts from a clinical data warehouse of the partners HealthCare system. Proc. AMIA Symp. 1174.

Murphy, S.N., Mendis, M.E., Berkowitz, D.A., Kohane, I., Chueh, H.C., 2006. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp. 2006, 1040.

Murphy, S.N., Morgan, M.M., Barnett, G.O., Chueh, H.C., 1999. Optimizing healthcare research data warehouse design through past COSTAR query analysis. Proc. AMIA Symp. 892–896.

Nardin, A., 2010. Le souci de l'humanisation de l'hôpital remonte à la fin des années 1920. La Croix.

Nielsen, J., 1993. Usability Engineering. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Nurseitov, N., Paulson, M., Reynolds, R., Izurieta, C., 2009. Comparison of JSON and XML data interchange formats: A case study, in: 22nd International Conference on Computer Applications in Industry and Engineering 2009, CAINE 2009. pp. 157–162.

Oracle Text [WWW Document], n.d. URL http://www.oracle.com/technetwork/testcontent/index-098492.html (accessed 4.6.17).

Ordonnance no 96-346 du 24 avril 1996 portant réforme de l'hospitalisation publique et privée, n.d.

Pakhomov, S., Pedersen, T., Chute, C.G., 2005. Abbreviation and Acronym Disambiguation in Clinical Discourse. AMIA. Annu. Symp. Proc. 2005, 589–593.

Park, A., Hartzler, A.L., Huh, J., McDonald, D.W., Pratt, W., 2015. Automatically Detecting Failures in Natural Language Processing Tools for Online Community Text. J. Med. Internet Res. 17. doi:10.2196/jmir.4612

Pasche, E., Chinali, M., Gobeill, J., Ruch, P., 2017. Development and Evaluation of a Case-Based Retrieval Service. Stud. Health Technol. Inform. 235, 186–190.

Patel, V.L., Kushniruk, A.W., 1998. Interface design for health care environments: the role of cognitive science. Proc. AMIA Symp. 29–37.

Plaisant, C., Lam, S., Shneiderman, B., Smith, M.S., Roseman, D., Marchand, G., Gillam, M., Feied, C., Handler, J., Rappaport, H., 2008. Searching Electronic Health Records for Temporal Patterns in Patient Histories: A Case Study with Microsoft Amalga. AMIA. Annu. Symp. Proc. 2008, 601–605.

Plovnick, R.M., Zeng, Q.T., 2004. Reformulation of consumer health queries with professional terminology: a pilot study. J. Med. Internet Res. 6, e27. doi:10.2196/jmir.6.3.e27

Prokosch, H.U., Ganslandt, T., 2009. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf. Med. 48, 38–44.

Puppala, M., He, T., Chen, S., Ogunti, R., Yu, X., Li, F., Jackson, R., Wong, S.T.C., 2015. METEOR: An Enterprise Health Informatics Environment to Support Evidence-Based Medicine. IEEE Trans. Biomed. Eng. 62, 2776–2786. doi:10.1109/TBME.2015.2450181

Raghavan, P., Chen, J.L., Fosler-Lussier, E., Lai, A.M., 2014. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci. 2014, 218–223.

Rijsbergen, C.J.V., 1979. Information Retrieval, 2nd ed. Butterworth-Heinemann, Newton, MA, USA.

Roden, D., Pulley, J., Basford, M., Bernard, G., Clayton, E., Balser, J., Masys, D., 2008. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clin. Pharmacol. Ther. 84, 362–369. doi:10.1038/clpt.2008.89

Rosenbloom, S.T., Denny, J.C., Xu, H., Lorenzi, N., Stead, W.W., Johnson, K.B., 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J. Am. Med. Inform. Assoc. JAMIA 18, 181–186. doi:10.1136/jamia.2010.007237

Ross, J., Tu, S., Carini, S., Sim, I., 2010. Analysis of Eligibility Criteria Complexity in Clinical Trials. Summit Transl. Bioinforma. 2010, 46–50.

Rubin, D.L., Desser, T.S., 2008. A data warehouse for integrating radiologic and pathologic data. J. Am. Coll. Radiol. JACR 5, 210–217. doi:10.1016/j.jacr.2007.09.004

Ruch, P., 2009. A medical informatics perspective on decision support: toward a unified research paradigm combining biological vs. clinical, empirical vs. legacy, and structured vs. unstructured data. Yearb. Med. Inform. 96–98.

Ruch, P., Baud, R., Geissbühler, A., 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artif. Intell. Med. 29, 169–184.

Ruch, P., Geissbühler, A., Gobeill, J., Lisacek, F., Tbahriti, I., Veuthey, A.-L., Aronson, A.R., 2007. Using discourse analysis to improve text categorization in MEDLINE. Stud. Health Technol. Inform. 129, 710–715.

Sadalage, P.J., Fowler, M., 2012. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Addison-Wesley.

Sager, N., Friedman, C., Chi, E., 1986. The analysis and processing of clinical narrative, in: In: Salamon R, Blum B, Jørgensen M, Editors. Presented at the Medinfo 86, Elsevier, Amsterdam (Holland), pp. 1101–5.

Salton, G., 1971. The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Salton, G., 1968. Automatic Information Organization and Retrieval. McGraw Hill Text.

Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G., 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. JAMIA 17, 507–513.

doi:10.1136/jamia.2009.001560

Scheufele, E., Aronzon, D., Coopersmith, R., McDuffie, M.T., Kapoor, M., Uhrich, C.A., Avitabile, J.E., Liu, J., Housman, D., Palchuk, M.B., 2014. tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. AMIA Summits Transl. Sci. Proc. 2014, 96–101.

Sengupta, D., Arora, P., Pant, S., Naik, P.K., 2013. Design of Dimensional Model for Clinical Data Storage and Analysis. Appl. Med. Inform. 32, 47–53.

Shin, S.-Y., Kim, W.S., Lee, J.-H., 2014. Characteristics Desired in Clinical Data Warehouse for Biomedical Research. Healthc. Inform. Res. 20, 109–116. doi:10.4258/hir.2014.20.2.109

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P.J., Elhadad, N., Johnson, S.B., Lai, A.M., 2014. A review of approaches to identifying patient phenotype cohorts using electronic health records. J. Am. Med. Inform. Assoc. JAMIA 21, 221–230. doi:10.1136/amiajnl-2013-001935

Shneiderman, B., 1997. Designing the User Interface: Strategies for Effective Human-Computer Interaction, 3rd ed. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M., Blike, G., 2016. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann. Intern. Med. 165, 753. doi:10.7326/M16-0961

Skeppstedt, M., 2011. Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. J. Biomed. Semant. 2 Suppl 3, S3. doi:10.1186/2041-1480-2-S3-S3

Smedley, D., Oellrich, A., Köhler, S., Ruef, B., Westerfield, M., Robinson, P., Lewis, S., Mungall, C., 2013. PhenoDigm: analyzing curated annotations to associate animal models with human diseases. Database J. Biol. Databases Curation 2013. doi:10.1093/database/bat025

South, B.R., Phansalkar, S., Swaminathan, A.D., Delisle, S., Perl, T., Samore, M.H., 2007. Adaptation of the NegEx algorithm to Veterans Affairs electronic text notes for detection of influenza-like illness (ILI). AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp. 2007, 1118.

Statistics - 2017AA Release [WWW Document], n.d. URL https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html (accessed 9.20.17).

Stead, W.W., Hammond, W.E., Straube, M.J., 1983. A chartless record--is it adequate? J. Med. Syst. 7, 103–109.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J., 2012. 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, EACL '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 102–107.

Stubbs, A., Kotfila, C., Xu, H., Uzuner, Ö., 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. J. Biomed. Inform. 58 Suppl, S67-77. doi:10.1016/j.jbi.2015.07.001

Stubbs, A., Uzuner, O., 2015. Annotating Risk Factors for Heart Disease in Clinical Narratives for Diabetic Patients. J. Biomed. Inform. 58, S78–S91. doi:10.1016/j.jbi.2015.05.009

Studnicki, J., Fisher, J.W., Eichelberger, C., Bridger, C., Angelon-Gaetz, K., Nelson, D., 2010. NC CATCH: Advancing Public Health Analytics. Online J. Public Health Inform. 2. doi:10.5210/ojphi.v2i3.3348

Sumathi, S., Esakkirajan, S., 2007. Fundamentals of Relational Database Management Systems. Springer Science & Business Media.

Sun, J., Wang, F., Hu, J., Edabollahi, S., 2012. Supervised Patient Similarity Measure of Heterogeneous Patient Records. SIGKDD Explor Newsl 14, 16–24. doi:10.1145/2408736.2408740

Sun, W., Rumshisky, A., Uzuner, O., 2013a. Temporal reasoning over clinical text: the state of the art. J. Am. Med. Inform. Assoc. JAMIA 20, 814–819. doi:10.1136/amiajnl-2013-001760

Sun, W., Rumshisky, A., Uzuner, O., 2013b. Annotating Temporal Information in Clinical Narratives. J. Biomed. Inform. 46. doi:10.1016/j.jbi.2013.07.004

Sun, W., Rumshisky, A., Uzuner, O., 2013c. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J. Am. Med. Inform. Assoc. JAMIA 20, 806–813. doi:10.1136/amiajnl-2013-001628

The Shire Rare Disease Impact Report (2013 – US and UK population), n.d.

Toll, E., 2012. The Cost of Technology. JAMA 307, 2497–2498. doi:10.1001/jama.2012.4946

Tomanek, K., Wermter, J., Hahn, U., 2007. A reappraisal of sentence and token splitting for life sciences documents. Stud. Health Technol. Inform. 129, 524–528.

Uzuner, Ö., 2009. Recognizing Obesity and Comorbidities in Sparse Data. J. Am. Med. Inform. Assoc. JAMIA 16, 561–570. doi:10.1197/jamia.M3115

Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., South, B.R., 2012. Evaluating the state of the art in coreference resolution for electronic medical records. J. Am. Med. Inform. Assoc. JAMIA 19, 786–791. doi:10.1136/amiajnl-2011-000784

Uzuner, Ö., Goldstein, I., Luo, Y., Kohane, I., 2008. Identifying Patient Smoking Status from Medical Discharge Records. J. Am. Med. Inform. Assoc. JAMIA 15, 14–24. doi:10.1197/jamia.M2408

Vallati, M., Gatta, R., De Bari, B., Magrini, S.M., 2013. Clinical similarities: an innovative approach for supporting medical decisions. Stud. Health Technol. Inform. 192, 1114.

Vuokko, R., Mäkelä-Bengs, P., Hyppönen, H., Lindqvist, M., Doupi, P., 2017. Impacts of structuring the electronic health record: Results of a systematic literature review from the perspective of secondary use of patient data. Int. J. Med. Inf. 97, 293–303. doi:10.1016/j.ijmedinf.2016.10.004

Wang, S., Pandis, I., Wu, C., He, S., Johnson, D., Emam, I., Guitton, F., Guo, Y., 2014. High dimensional biological data retrieval optimization with NoSQL technology. BMC Genomics 15, S3. doi:10.1186/1471-2164-15-S8-S3

Weber, G.M., Murphy, S.N., McMurry, A.J., MacFadden, D., Nigrin, D.J., Churchill, S., Kohane, I.S.,

2009. The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. J. Am. Med. Inform. Assoc. JAMIA 16, 624–630. doi:10.1197/jamia.M3191

Wei, W.-Q., Denny, J.C., 2015. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med. 7, 41. doi:10.1186/s13073-015-0166-y

Weng, C., Bigger, J.T., Busacca, L., Wilcox, A., Getaneh, A., 2010. Comparing the effectiveness of a clinical registry and a clinical data warehouse for supporting clinical trial recruitment: a case study. AMIA Annu. Symp. Proc. AMIA Symp. 2010, 867–871.

Wisniewski, M.F., Kieszkowski, P., Zagorski, B.M., Trick, W.E., Sommers, M., Weinstein, R.A., 2003. Development of a clinical data warehouse for hospital infection control. J. Am. Med. Inform. Assoc. JAMIA 10, 454–462. doi:10.1197/jamia.M1299

Wu, A.S., Do, B.H., Kim, J., Rubin, D.L., 2011. Evaluation of Negation and Uncertainty Detection and its Impact on Precision and Recall in Search. J. Digit. Imaging 24, 234–242. doi:10.1007/s10278-009-9250-4

Wu, S., Miller, T., Masanz, J., Coarr, M., Halgrim, S., Carrell, D., Clark, C., 2014. Negation's not solved: generalizability versus optimizability in clinical natural language processing. PloS One 9, e112774. doi:10.1371/journal.pone.0112774

Wu, S.T., Liu, H., Li, D., Tao, C., Musen, M.A., Chute, C.G., Shah, N.H., 2012. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. J. Am. Med. Inform. Assoc. JAMIA 19, e149–e156. doi:10.1136/amiajnl-2011-000744

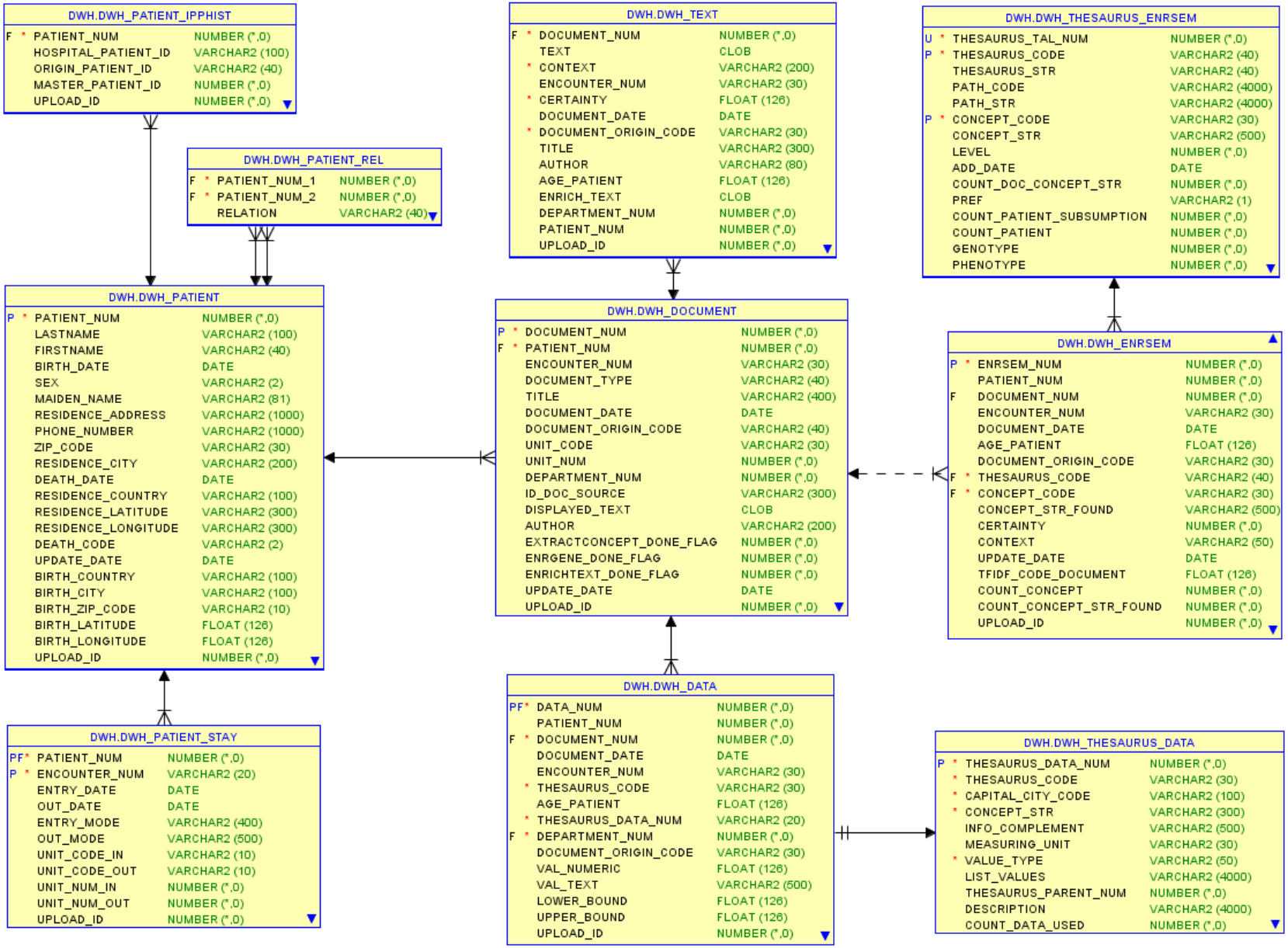
Wu, Y., Xu, J., Jiang, M., Zhang, Y., Xu, H., 2015. A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. AMIA. Annu. Symp. Proc. 2015, 1326–1333.

Xu, H., Stenner, S.P., Doan, S., Johnson, K.B., Waitman, L.R., Denny, J.C., 2010. MedEx: a medication information extraction system for clinical narratives. J. Am. Med. Inform. Assoc. 17, 19–24. doi:10.1197/jamia.M3378

Yamamoto, K., Sumi, E., Yamazaki, T., Asai, K., Yamori, M., Teramukai, S., Bessho, K., Yokode, M., Fukushima, M., 2012. A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research. BMJ Open 2. doi:10.1136/bmjopen-2012-001622

Zhang, G.-Q., Siegler, T., Saxman, P., Sandberg, N., Mueller, R., Johnson, N., Hunscher, D., Arabandi, S., 2010. VISAGE: A Query Interface for Clinical Research. Summit Transl. Bioinforma. 2010, 76–80.

Annexe 1 : Schéma simplifié du modèle de données de Dr Warehouse



Annexe 2

Format XML de stockage d'une requête. Le bloc atomic_query est répété autant de fois qu'il y a de requêtes atomiques

```
<query>
      <atomic_query>
            <textual_query>
                  <text></text>
                  <synonym_expansion></synonym_expansion>
                  <context></context>
                  <certainty></certainty>
            </textual_query>
            <str_structured_query>
                  <thesaurus></thesaurus>
                  <thesaurus_data_num></thesaurus_data_num>
                  <thesaurus_concept_code></thesaurus_concept_code>
                  <out_bound></out_bound>
                  <operator></operator>
                  <value></value>
                  <min_value></min_value>
                  <max_value></max_value>
                  <value_sup_n_x_bound_upper></value_sup_n_x_bound_upper>
                  <value_inf_n_x_bound_lower></value_inf_n_x_bound_lower>
                  <list_value_available></list_value_available>
            </str_structured_query>
            <exclude></exclude>
            <document_origin_code></document_origin_code>
            <document_date_start></document_date_start>
            <document_date_end></document_date_end>
            <period_document></period_document>
            <document_ageyear_start></document_ageyear_start>
            <document_ageyear_end></document_ageyear_end>
            <document_agemonth_start></document_agemonth_start>
            <document_agemonth_end></document_agemonth_end>
            <filter_num></filter_num>
            <datamart_text_num></datamart_text_num>
            <hospital_department_list></hospital_department_list>
            <count_result></count_result>
            <query_type></query_type>
      </atomic_query>
      <datamart_num></datamart_num>
      <sex></sex>
      <age_start></age_start>
      <age_end></age_end>
      <alive_death></alive_death>
      <age_death_start></age_death_start>
      <age_death_end></age_death_end>
      <first_stay_date_start></first_stay_date_start>
      <first_stay_date_end></first_stay_date_end>
```

```
      <minimum_period_folloup></minimum_period_folloup>
      <list_excluded_cohort></list_excluded_cohort>
</query>
```