

Un entrepôt de données pour un système de santé apprenant

Et pour les maladies rares !

Nicolas Garcelon

Responsable de la plateforme data science, Institut Imagine

Déclaration d'intérêts

Co-fondateur en 2017 d'une spin-off de l'institut imagine : codoc

Un contexte de recherche translationnelle



Activité clinique
Bases de données
hospitalières



600 lits
- 400 lits pédiatriques
- 200 lits adultes



34 centres de Référence pour les maladies rares
25 centres de compétences



Recherche génétique
Bases de données
recherches



30 laboratoires
12 plateformes

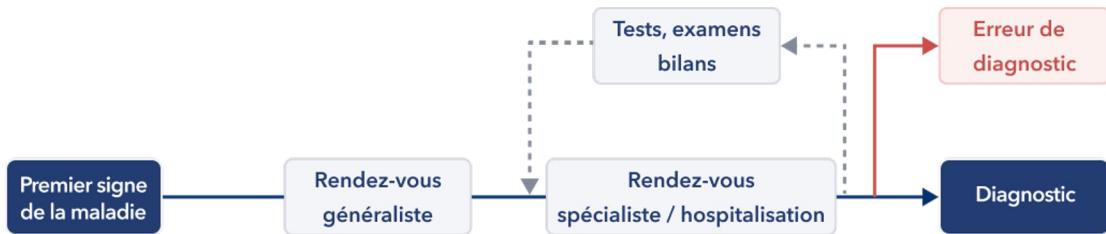
Le contexte des maladies rares

+7000 maladies rares connues
(250 découvertes chaque année)

300 millions de personnes atteintes dans le monde*
1 personne sur 20 en France

2/3 des maladies rares sont graves et génèrent un handicap**

Un patient peut attendre en moyenne
5 ans pour être diagnostiqué :
errance diagnostique et thérapeutique



La lutte contre l'errance diagnostique et thérapeutique, l'amélioration du parcours de soins et l'accès à l'information constituent les priorités du plan national maladies rares.

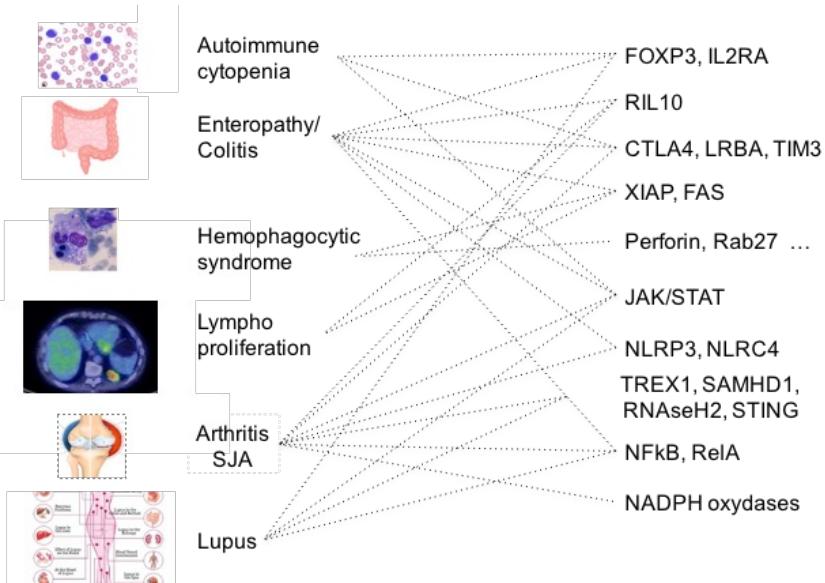
*Nguengang Wakap, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 28, 165–173 (2020).

**https://www.has-sante.fr/upload/docs/application/pdf/avis_ald_rapport.pdf

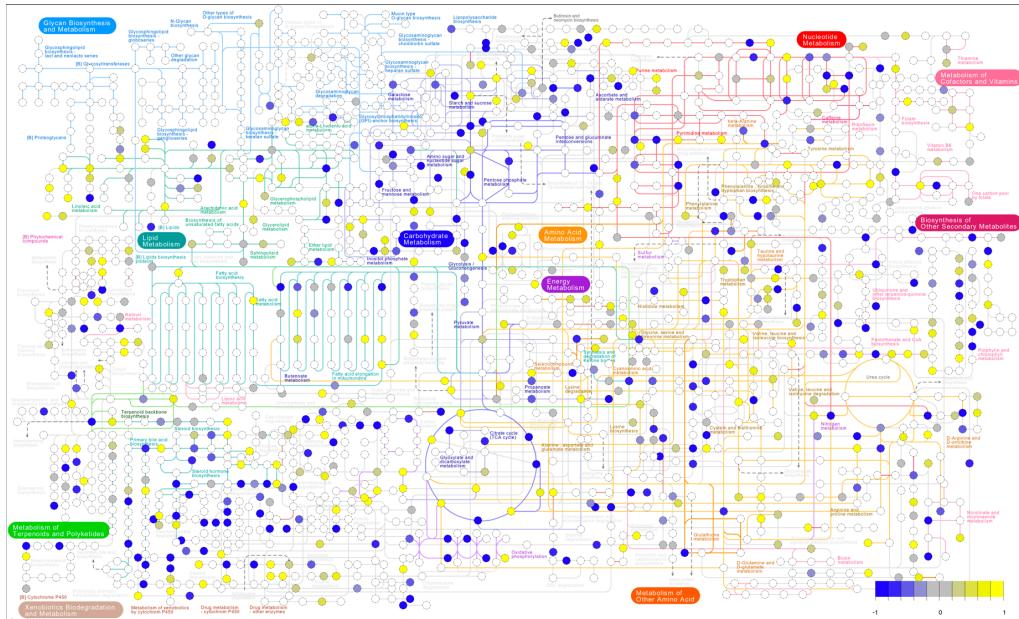
Le contexte génétique

Complexité de l'expression génétique :

- Pléiotropie : plusieurs actions pour un gène
- Polygénie : plusieurs gènes pour la même action



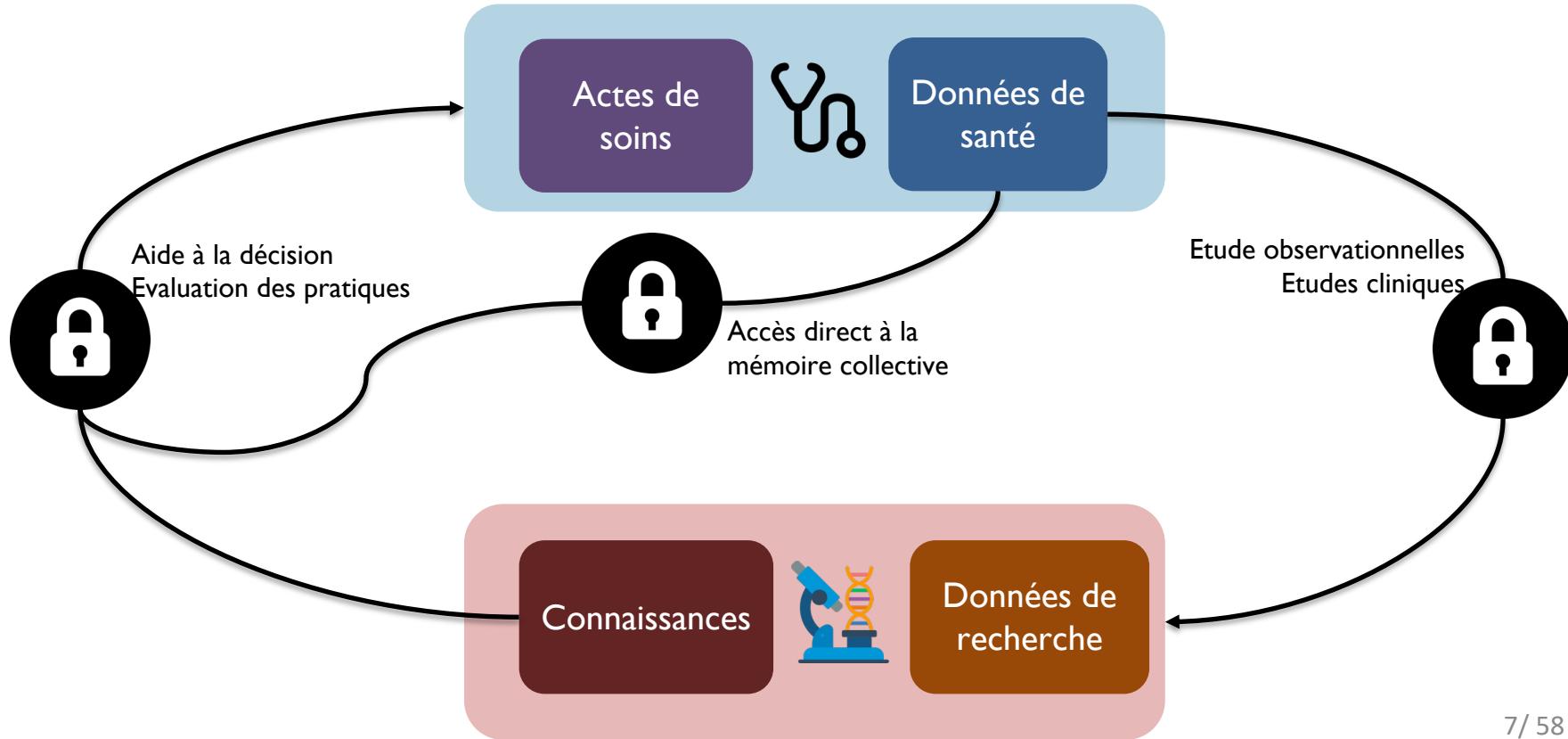
© Illustration projet ATRAction – Rieux Laucat



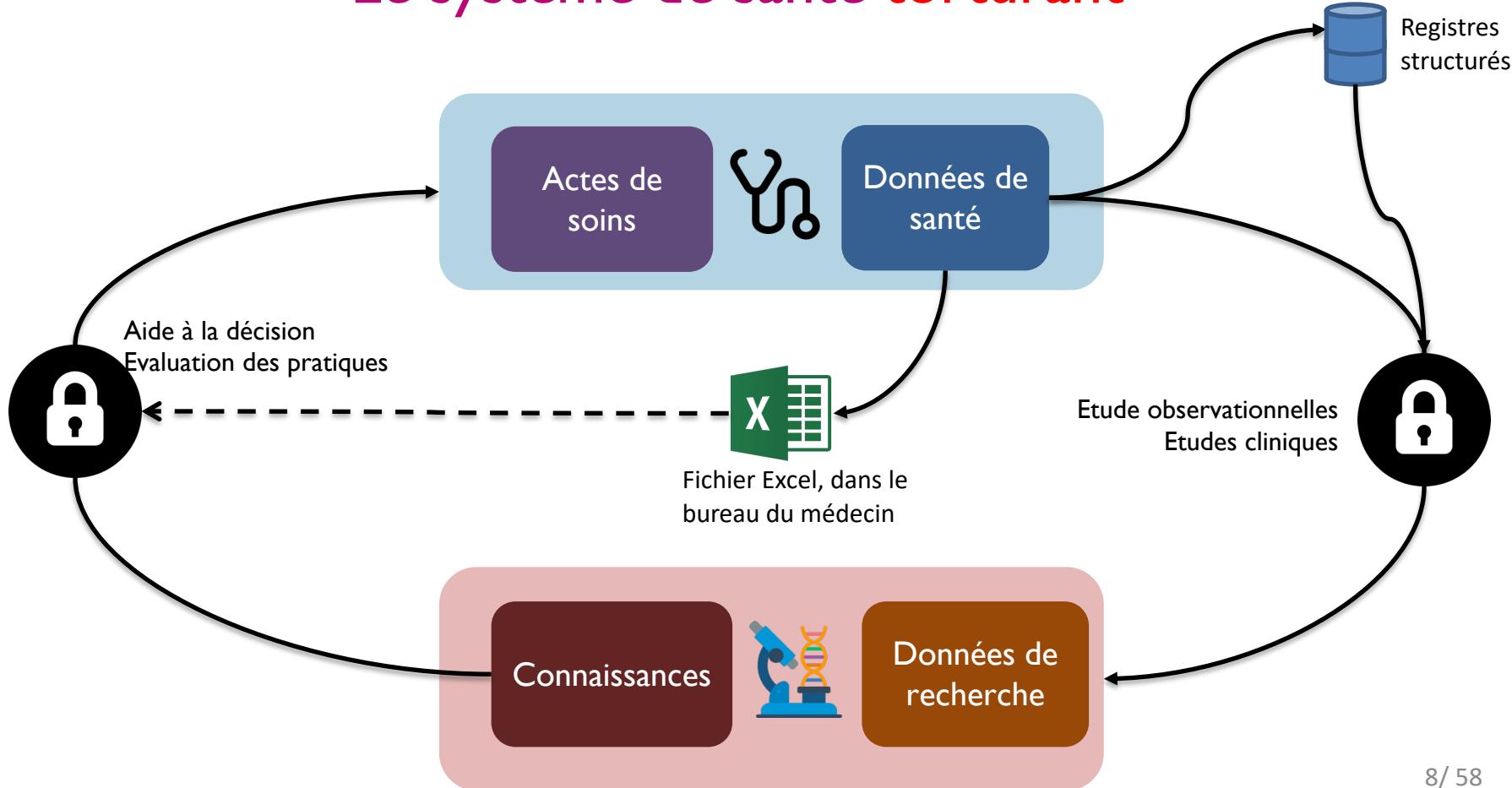
Des données multiples

- **Médicales issues de la cliniques (dossier patient)**
 - ✓ Comptes rendus hospitalisations, consultations
 - ✓ Examens Biologiques
 - ✓ Examens complémentaires
 - ✓ Dossier infirmier
 - ✓ Anatomopathologie
 - ✓ Imagerie
 - ✓ Parcours de soin
 - ✓ Traitements
 - ✓ Etc.
- **Environnementales (pollution, météorologie, etc.)**
- **Moléculaires issues de la recherche**
 - ✓ Génomique
 - ✓ Exomique
 - ✓ Transcriptomique
 - ✓ Epigénétique
 - ✓ Protéomique
 - ✓ Métabolomique
 - ✓ Météagénomique
- **Bases de connaissance : ontologie de phénotype, pathways moléculaires, lien de causalité, modèles biologiques, modèles animaux**

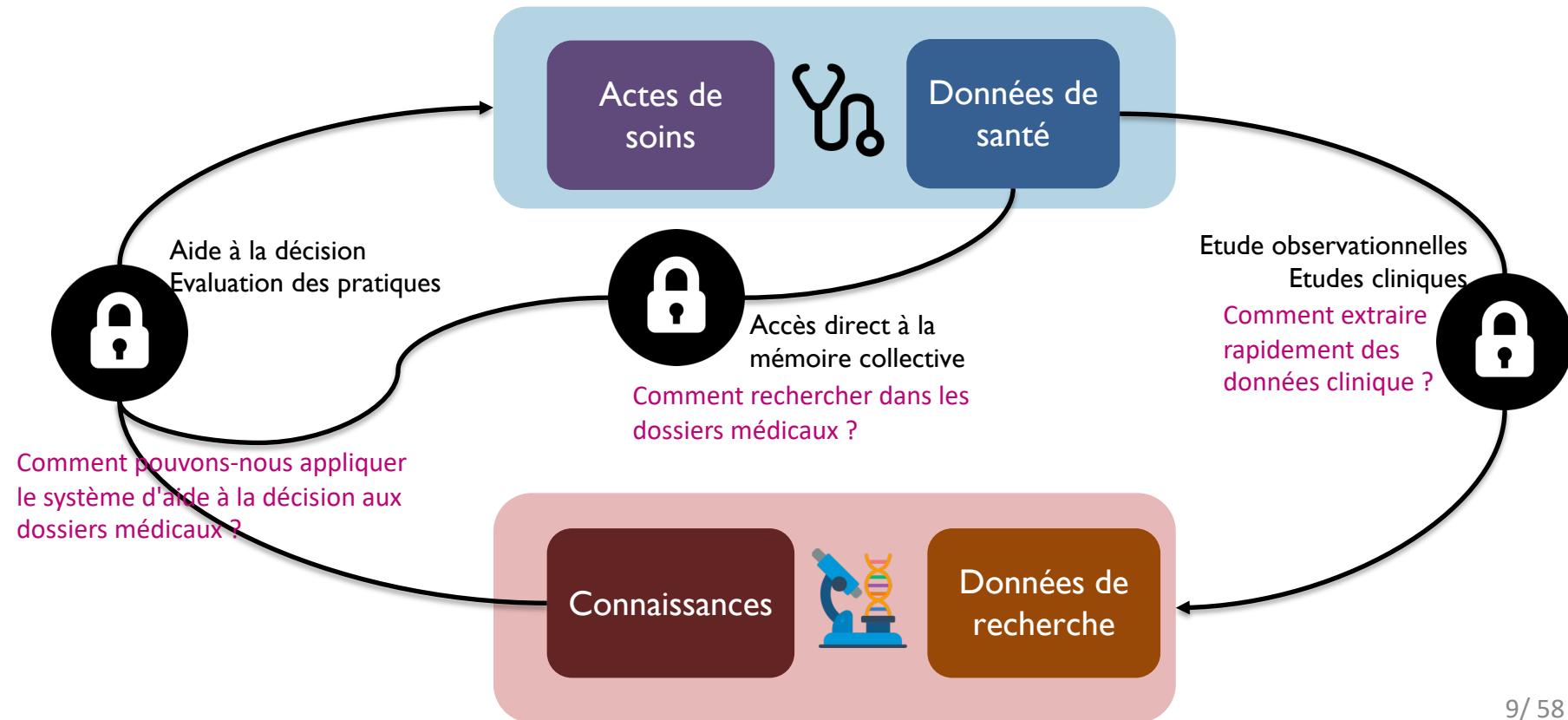
Le système de santé apprenant



Le système de santé torturant

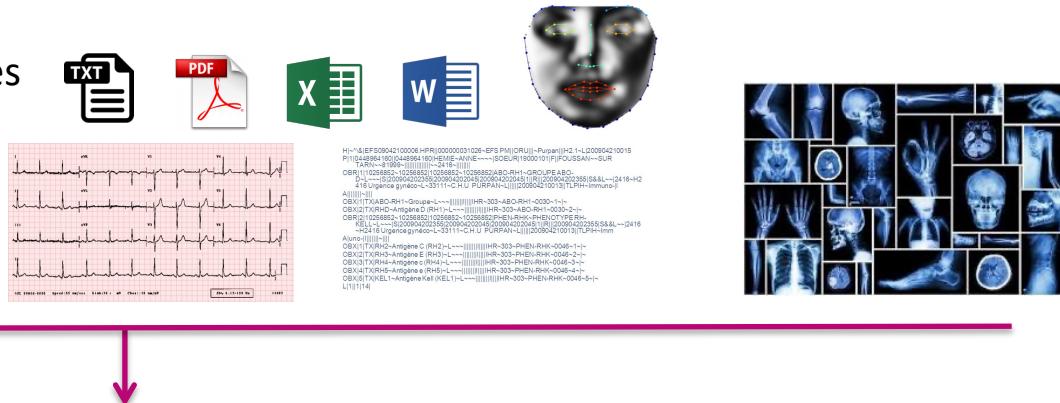


Le système de santé apprenant



Principal verrou à la réutilisation : l'accessibilité aux données

- Format propriétaire / accès aux données
 - Hétérogénéité des formats
 - Cloisonnement des données
 - Volume des données



Entrepôt de données

Idées clés :

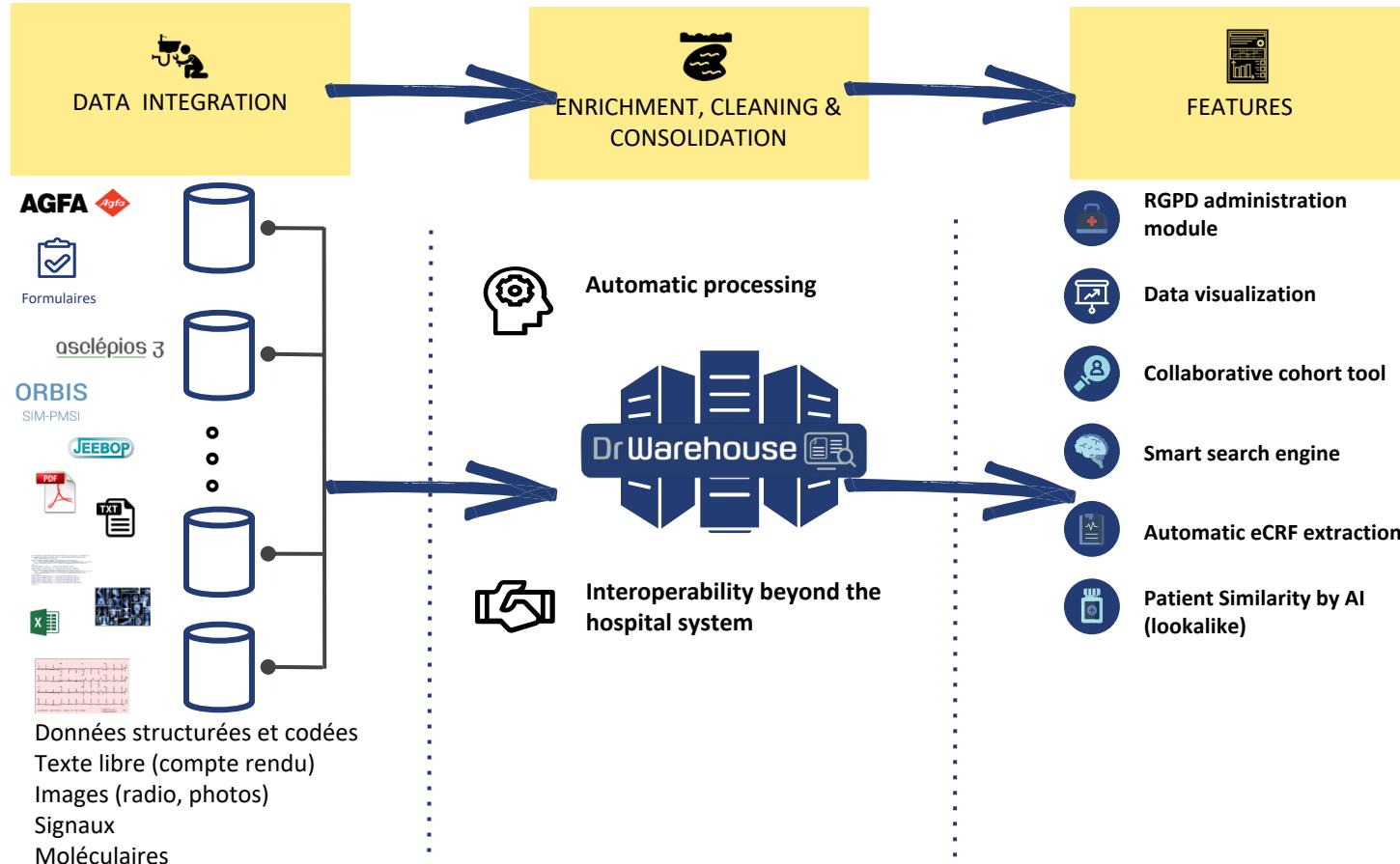
Intégrer les données multimodales hospitalières et les rendre exploitables par les

Médecins* et data scientist**

⇒ *Médecins : une interface intuitive, peu de clics, efficacité du logiciel

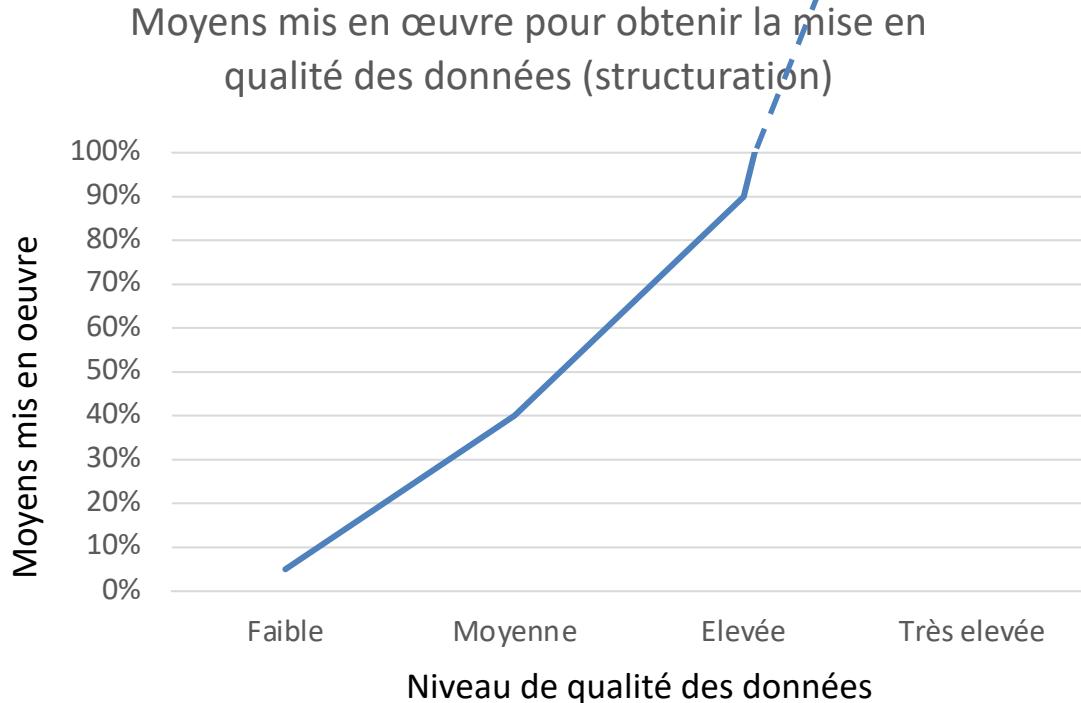
⇒ **Data scientist : peu importe ☺

Intégration de toutes les données hospitalières



Deuxième verrou à la réutilisation : la mise en qualité des données

Le niveau de qualité dépend de l'objectif et des moyens à mettre en œuvre pour y arriver. Cela nécessite de faire des choix en lien avec le contexte de réutilisation des données.



Attention : Les données brutes ne sont pas des données sales pour le soin ! Elles sont sales pour une réutilisation dans un contexte d'étude épidémiologique etc.

Graphique complètement inventé sans aucune valeur scientifique dont l'objectif est d'illustrer mon propos

Deuxième verrou à la réutilisation : la mise en qualité des données

Le niveau de qualité dépend de l'objectif et du niveau d'intervention humaine pour y arriver.
Cela nécessite de faire des choix en lien avec le contexte de réutilisation des données.
Lorsque la donnée est transformée puis exportée, le contexte est perdu.

Idées clés :

- Conserver les données brutes, pour garder le contexte. Néanmoins, améliorer la qualité sur les données administratives (Identitovigilance etc.)
- Transformation automatique des données pour faciliter la génération d'hypothèse
 - => Attention à la qualité : non validées manuellement
 - => Erreur dans la temporalité : exemple copier coller d'un compte-rendu à l'autre
 - => Possibles Faux positifs, Faux négatifs : pouvoir les quantifier
- Développer des outils pour faciliter l'extraction et le contrôle qualité pour une analyse.
- Création de métadonnées précises pour décrire les données extraites

Intégration de données hétérogènes, structurées, textuelles

- Format propriétaire / accès aux données
- Hétérogénéité des formats
- Cloisonnement des données
- Volume des données
- Données structurées
- Données textuelles



Entrepôt de données hospitalier

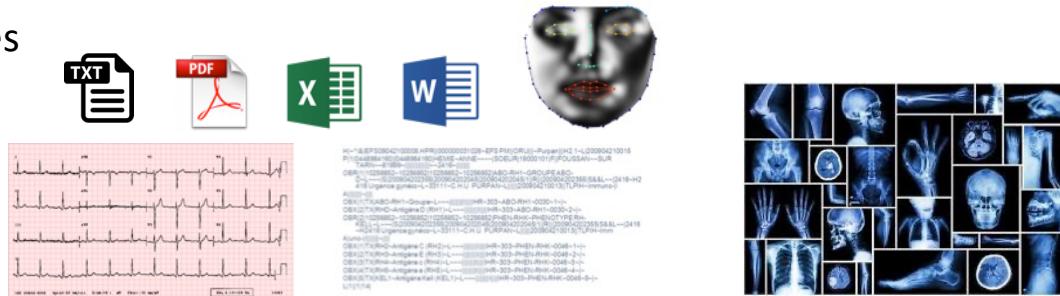
Accès à la mémoire

Fouille de données

Accélérer la recherche translationnelle

Intégration de données hétérogènes, structurées, textuelles

- Format propriétaire / accès aux données
- Hétérogénéité des formats
- Cloisonnement des données
- Volume des données
- Données structurées
- Données textuelles



Entrepôt de données hospitalier

Accès à la mémoire

Moteur de recherche

- ✓ Recruter des patients
- ✓ Tester d'hypothèse

Fouille et extraction de données

Inférence sur les données

- ✓ Connecter les données à des connaissances
- ✓ Générer de nouvelles hypothèses

Accélérer la recherche translationnelle

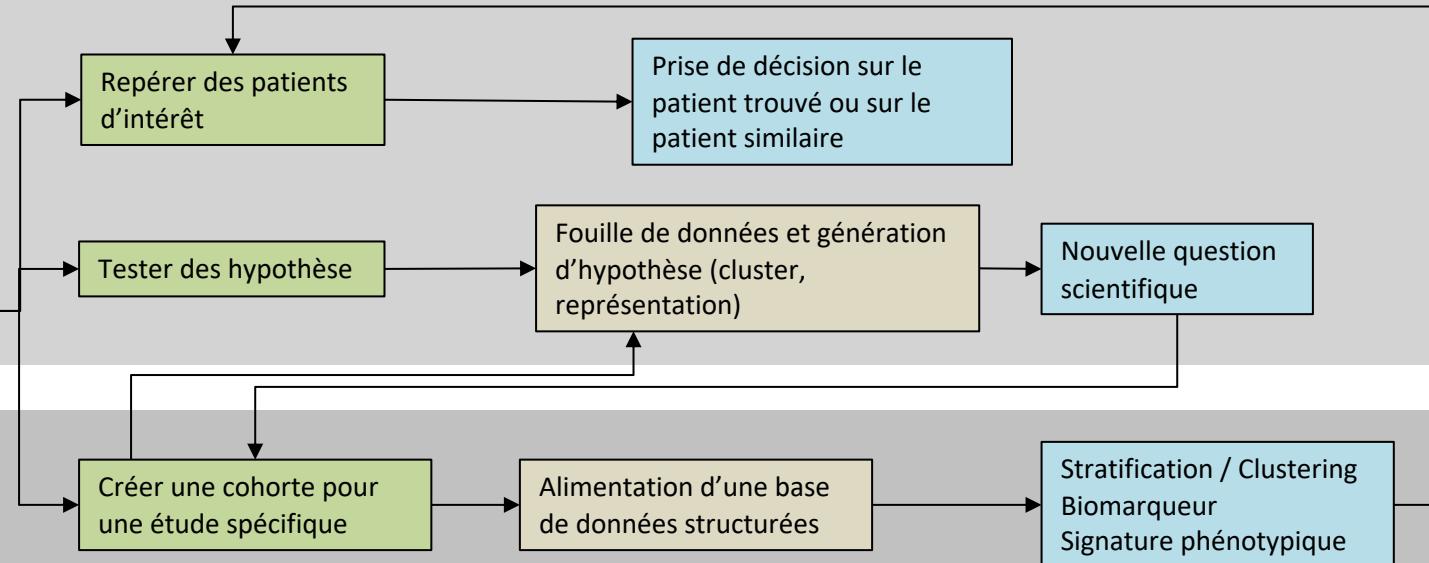
Fonctionnalités centrées patients

- ✓ Faciliter l'alimentation d'un data set
- ✓ Faciliter l'application d'outils d'aide à la décision

Dans le service du médecin : périmètre de soin

Dans un usage quotidien !

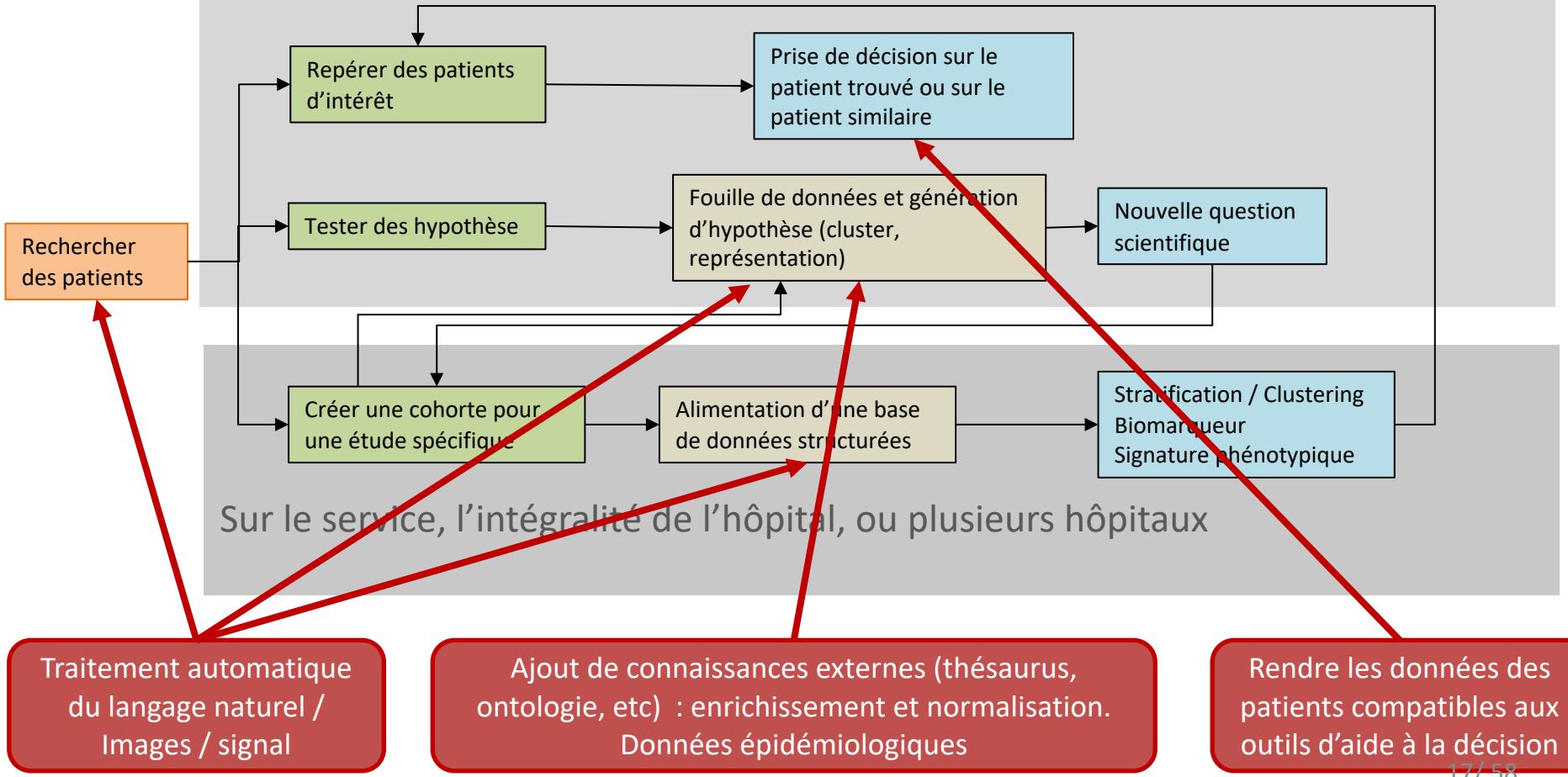
Rechercher
des patients



Sur le service, l'intégralité de l'hôpital, ou plusieurs hôpitaux



Dans le service du médecin : périmètre de soin



Dr Warehouse à Necker

850 000 patients

9 millions documents

70 millions données structurées

33 sources of data

From 1996 to 2023

Comptes rendus de consultation

Comptes rendus d'hospitalitReports

Comptes rendus d'imagerie

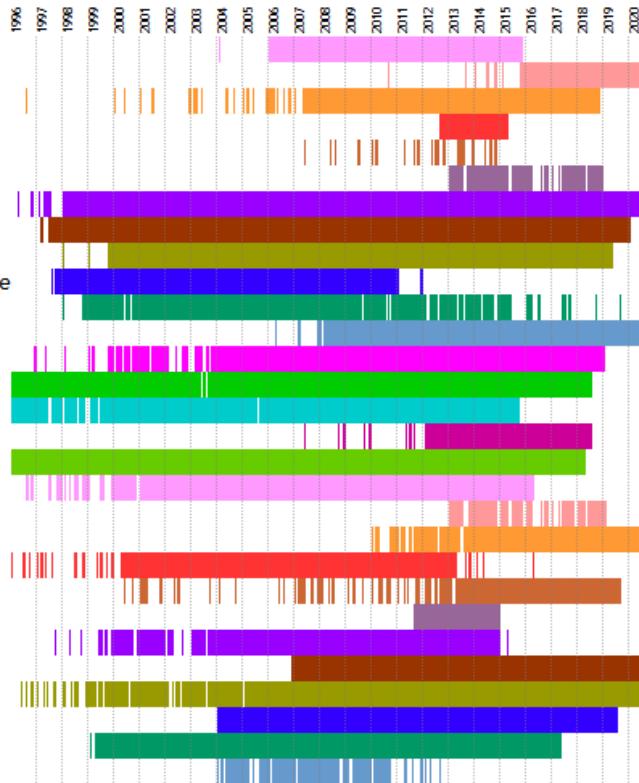
Comptes rendus d'anapath

Comptes rendus de foetopathologie

Comptes rendus de génétiques

Résultats biologiques

Anapath APIX
Anapath DIAMIC
ASTRAIA
Base IRM NB
Base MCD NBB
Base NPH
Cardiologie Commun
Cathétérisme Cardiaque
Chirurgie cardio vasculaire
Chirurgie orthopédique pédiatrique
Chirurgie Viscérale Commun
CR RCP
Diammg
Echographie CARDIAQUE
Foetopath
MEDIWEB source inconnue
Néphro Adulte Commun
Néphro Péd Commun
NPH Commentaire
ORBIS
ORL Troussseau Commun
PACS
PMSI-CIM10
RADOS
STARE
SUSIE
Transplantation rénale
Traumatologie Commun CMI
Traumatologie Commun CRH



Moteur de recherche sur les données brutes

Multimodalité :

- Données textuelles
- Données structurées
- Mouvements des patients
- Critères temporels
- Données démographiques

Accélérer

- Recrutement de patients
- Test d'hypothèses
- Création de cohortes

The screenshot shows the Dr. Care web application interface. At the top, there's a navigation bar with links for Accueil, Recherche, Cohortes, Tableau de bord, and Admin. On the far right, there's a search bar labeled "Rechercher un patient" and some user statistics (78 notifications, 811 messages). The main area is divided into two sections: "Recherche" on the left and "Résultats" on the right.

In the "Recherche" section, there's a search bar with the query "rett syndrome" OR FOXG1. Below it, there are filters for "Population" (4 651 patients) and "Critère Textuel" (78 patients), both of which are expanded. There are also checkboxes for "Étendre aux synonymes" and "Filtres avancés". A "Rechercher" button is at the bottom of this section.

The "Résultats" section has a header "Récapitulatif de la recherche" with a note about excluding negations. It includes tabs for "Dossiers patients" (selected), "Démographie", "Biologie", and "PMSI". Below this, there's a button "Alimenter une cohorte".

The main results list contains three entries:

- GILLETTE Rebecca**, F, 39, Née le 10/08/1959 - 63 ans. Associated documents: *Brain & development, le 01/11/2008, par Dr Harada Koto...*, *Proceedings of the National Academy of Sciences of the...*, *Advances in experimental medicine and biology, le...*. A "Afficher plus" button is shown.
- KNOWLES James**, H, 40, Né le 29/01/1954 - 69 ans. Associated documents: *Stem cell investigation, le 14/04/2012, par Dr Gomathi...*, *Developmental neurorehabilitation, le 15/06/2007, par D...*, *Epilepsia, le 30/03/2007, par Dr Lim Zhan - Pubmed...*. A "Afficher plus" button is shown.
- LEE David**, H, 41. Associated documents: *Current biology : CB, le 14/04/2012, par Dr Franco Luis M ...*, *Stem cell research, le 14/04/2012, par Dr Hunihan Lisa -...*. A "Afficher plus" button is shown.

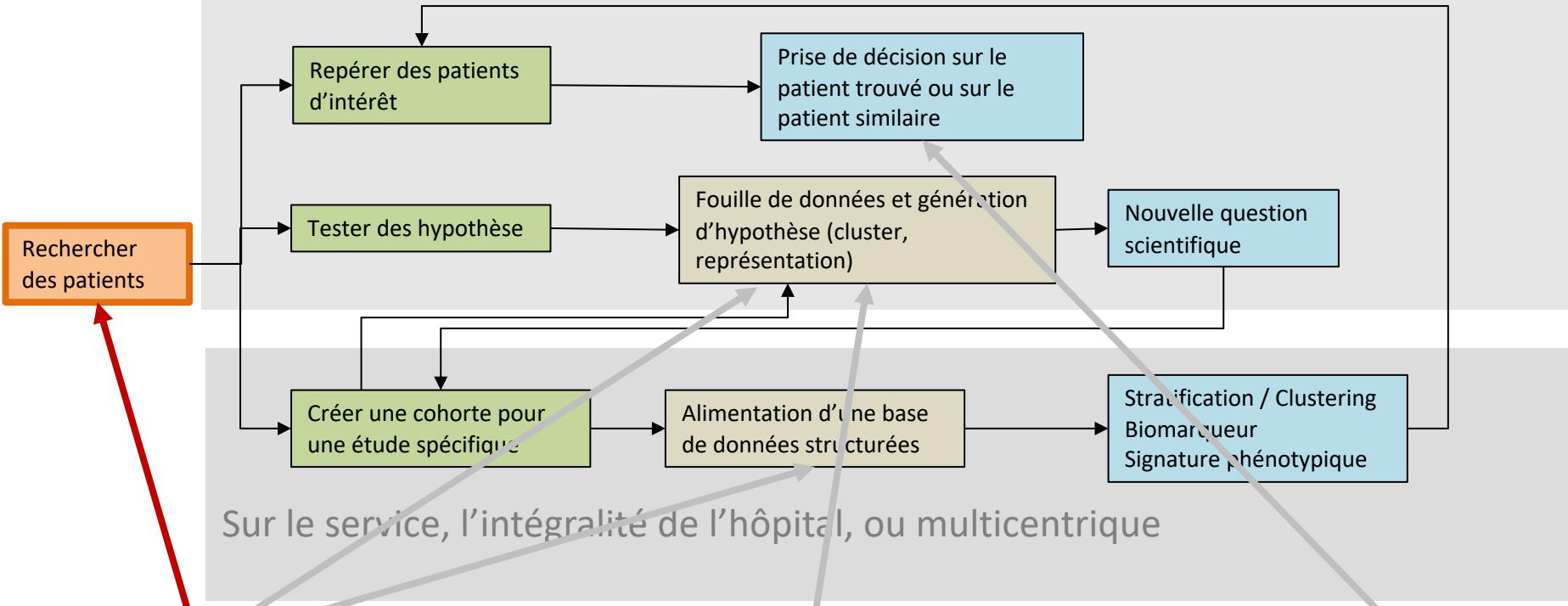
At the bottom of the results page, there are navigation buttons (1, 2, 3, 4, 5, >) and a "78 résultat(s) / 20 par page" link. To the right of the results, there's a sidebar with a summary of the first result and a detailed view of the "Brain & development" document.

Verrous liés aux données textuelles

- Bruit généré : négations, antécédents familiaux (*le patient n'a pas d'insuffisance rénale. La mère a un diabète de type 2*) => nécessité de développer une stratégie de traitement automatique du langage naturel

® Capture Suite codoc

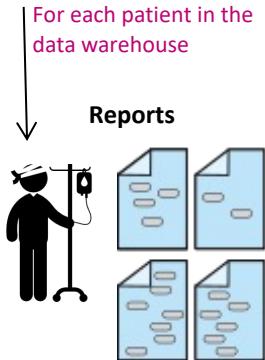
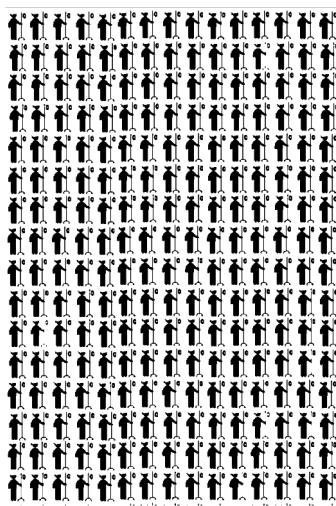
Dans le service du médecin



Traitement automatique
du langage naturel /
Images / signal

Ajout de connaissances externes (thésaurus,
ontologie, etc) : enrichissement et normalisation.
Données épidémiologiques

Rendre les données des
patients compatibles aux
outils d'aide à la décision
20/ 58



Améliorer le moteur de recherche pas une approche à base de règles

TraITEMENT du langage naturel pour détECTER

- Négation *Nous pouvons exclure le diagnostic de Syndrome de Rett*
- Antécédents familiaux *Le père est décédée d'un cancer du pancréas*
- Hypothèse *Suspicion de lupus. Recherche de mutation sur LRBA1*

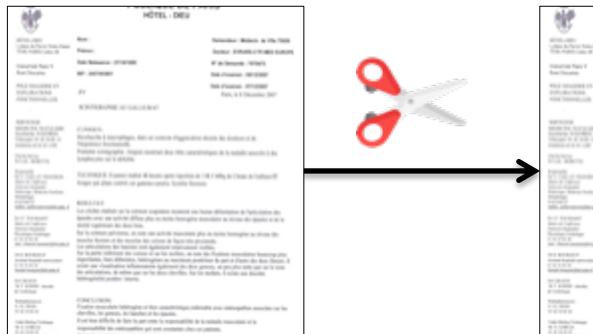
- 1- Découpage en phrase
- 2- DéteCTIONS d'éléments en lien avec la famille
- 3- Découpage en syntagme
- 4- Détection des négations

Déterminer l'expression de la négation et des antécédents familiaux

Objectif : conserver le texte libre pour le moteur de recherche

Notre approche est de découper le texte en syntagmes et les classer en fonction du contexte et de la certitude à partir de règles : présence de mots et d'expressions exprimant la négation ou l'antécédent familial.

Compte rendu



Classification

Règles :
Pas de
Absence de
Etc.

Cousin
Sœur
Etc.

| Texte | Certitude | Contexte |
|--|-------------|----------|
| <p>Malheureusement, je ne pourrai pas vous aider à ce sujet.</p> <p>Malheureusement, je ne pourrai pas vous aider à ce sujet.</p> <p>Malheureusement, je ne pourrai pas vous aider à ce sujet.</p> | Affirmation | Patient |
| <p>Malheureusement, je ne pourrai pas vous aider à ce sujet.</p> <p>Malheureusement, je ne pourrai pas vous aider à ce sujet.</p> | Négation | Patient |
| <p>Malheureusement, je ne pourrai pas vous aider à ce sujet.</p> <p>Malheureusement, je ne pourrai pas vous aider à ce sujet.</p> | Affirmation | Famille |
| | Négation | Famille |

Déetecter l'expression de la négation et des antécédents familiaux

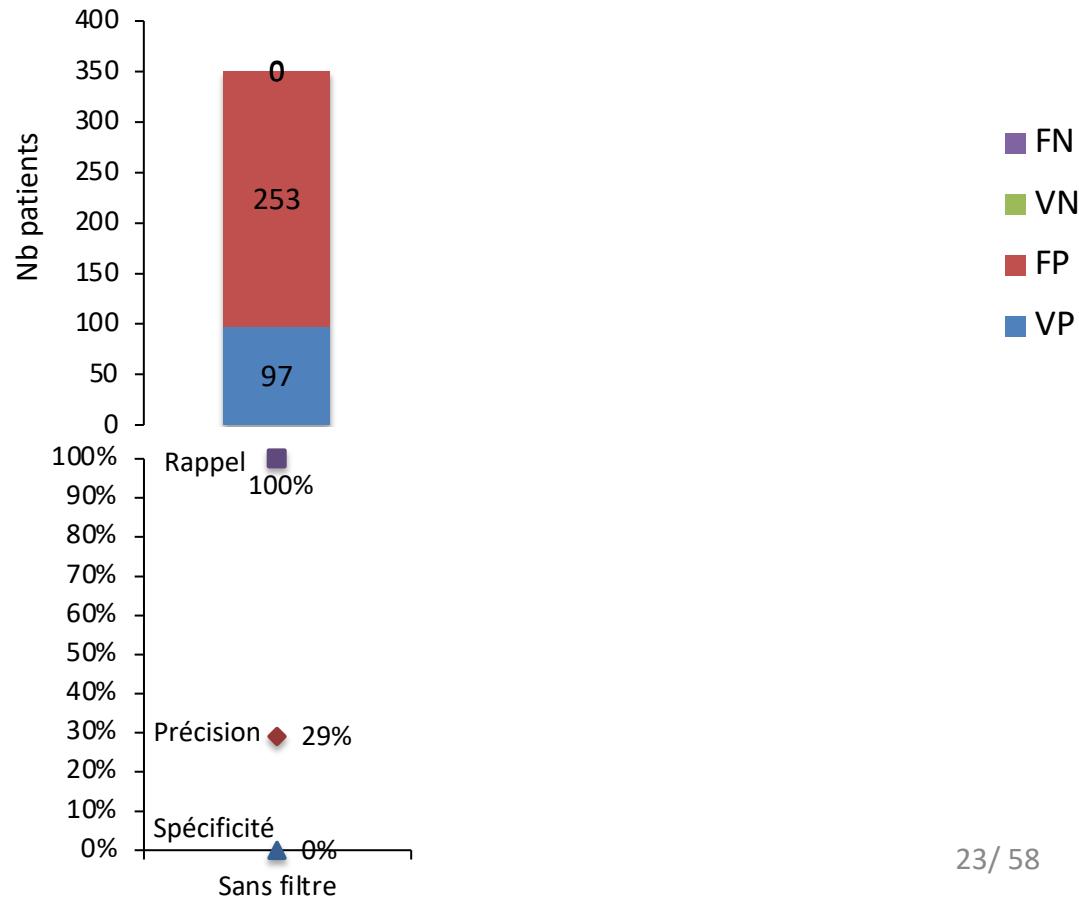
Résultats

Evaluation sur 3 requêtes :

« lupus and diarrhée » 145 patients (262 documents)

« crohn and diabète » 173 patients (269 documents)

« NPHP1 » 32 patients (95 documents)



Déetecter l'expression de la négation et des antécédents familiaux

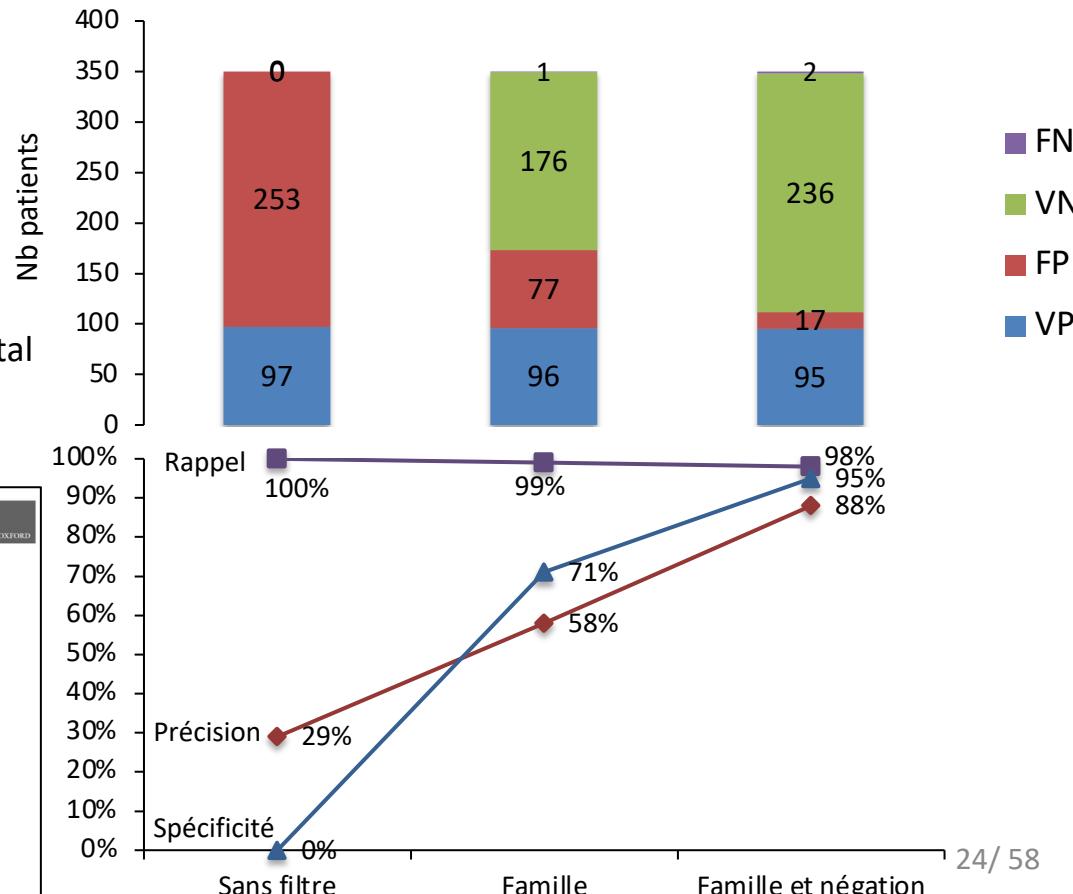
Résultats

Evaluation sur 3 requêtes :

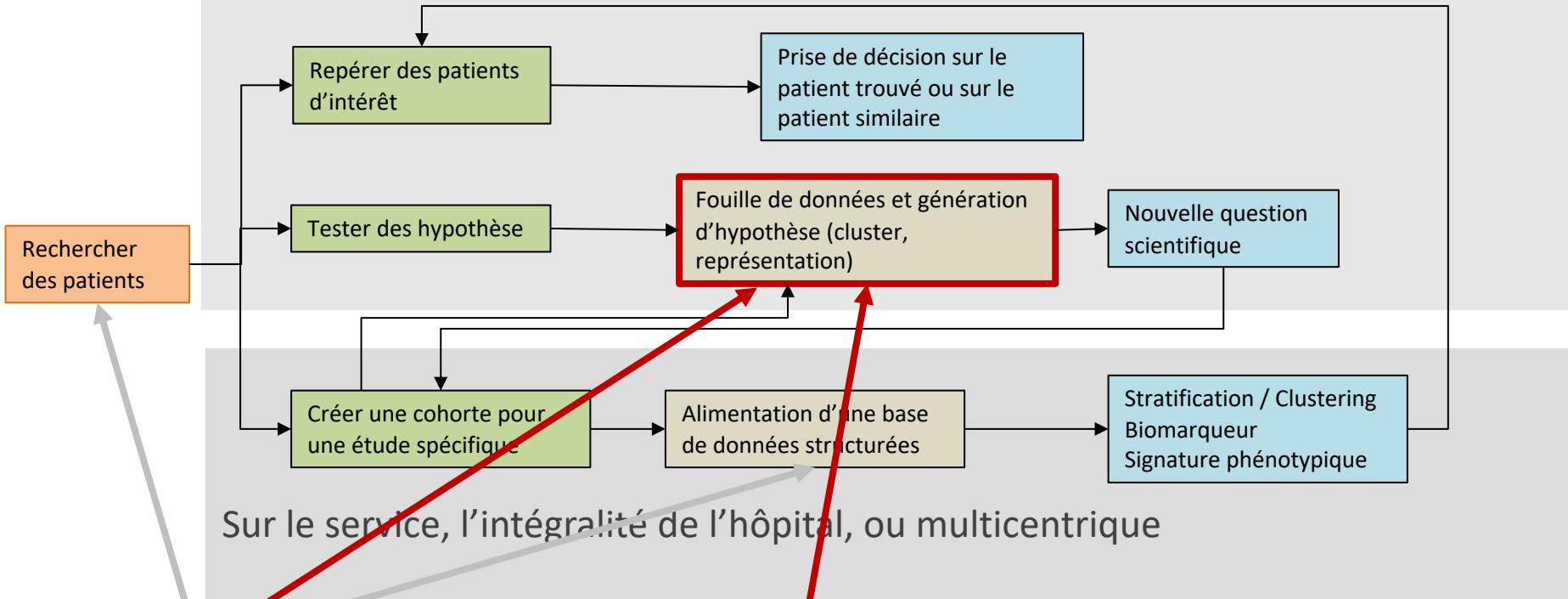
- « lupus and diarrhée » 145 patients (262 documents)
- « crohn and diabète » 173 patients (269 documents)
- « NPHP1 » 32 patients (95 documents)

Sur les 500 000 patients, le moteur renvoie un total de 102 patients dont :

- 95 vrais positifs
- 17 faux positifs
- Rappel : 0.98
- Précision : 0.88
- Spécificité : 0.95



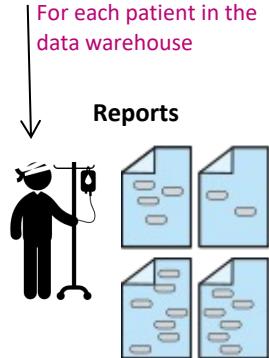
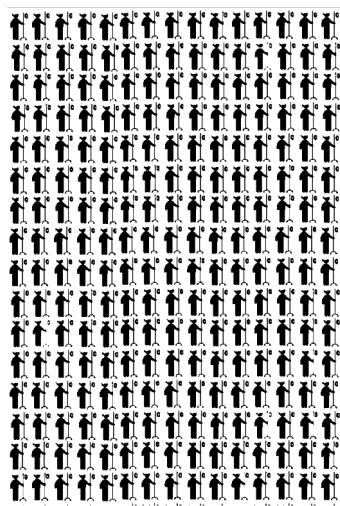
Dans le service du médecin



Traitements automatiques
du langage naturel /
Images / signal

Ajout de connaissances externes (thèses, ontologie, etc) : enrichissement et normalisation.
Données épidémiologiques

Rendre les données des patients compatibles aux outils d'aide à la décision
25/ 58



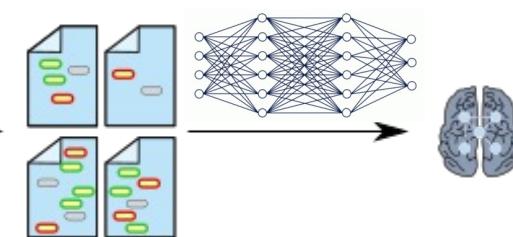
L'intelligence artificielle pour transformer le dossier patient en données structurées

Traitements du langage naturel pour détecter

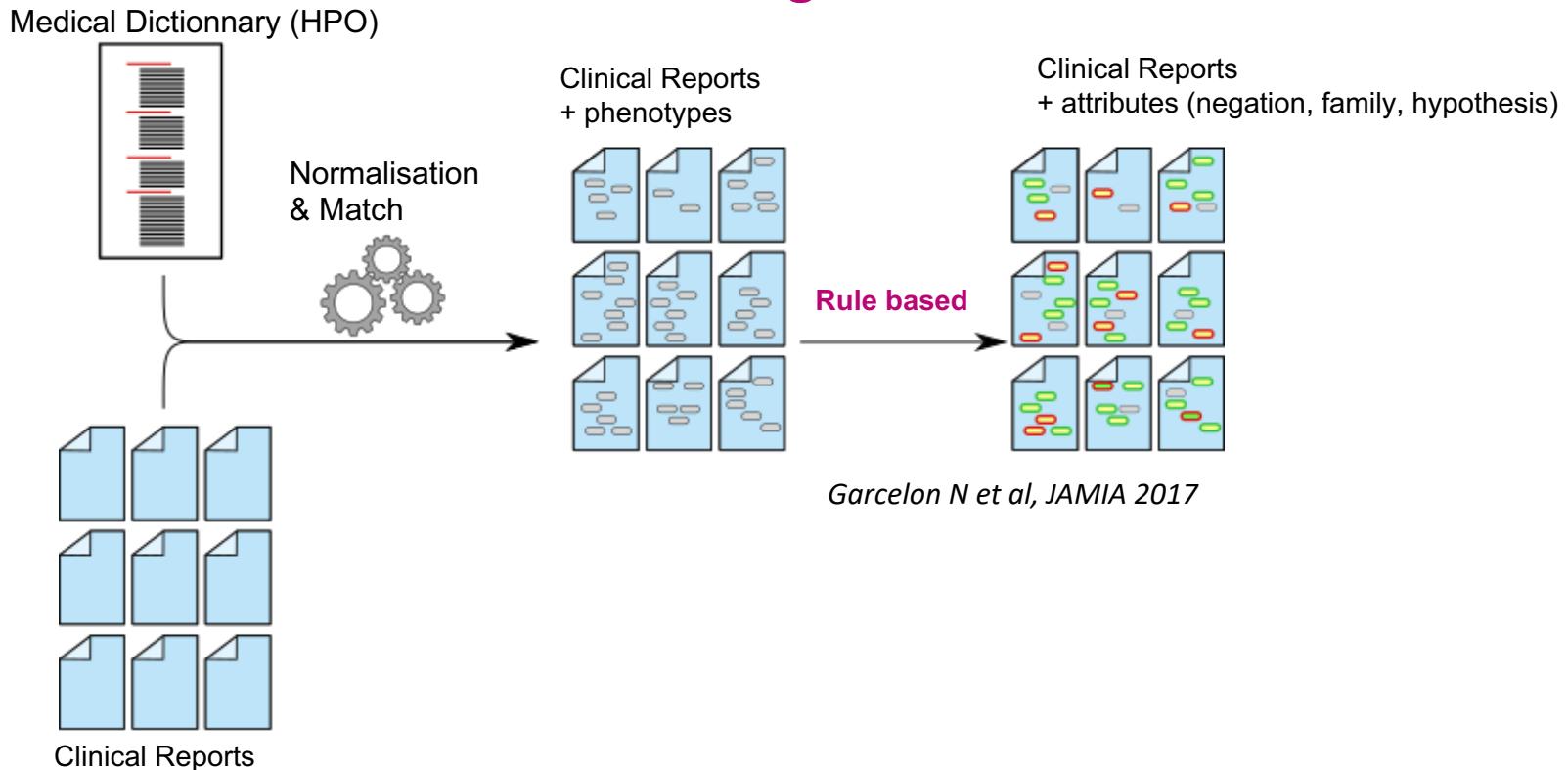
- Négation *Nous pouvons exclure le diagnostic de Syndrome de Rett*
- Antécédents familiaux *Le père est décédée d'un cancer du pancréas*
- Hypothèse *Suspicion de lupus. Recherche de mutation sur LRBA1*
- Extraire des phénotypes et des diagnostics

Extraire informations depuis des photos (landmarks, morphométrie)

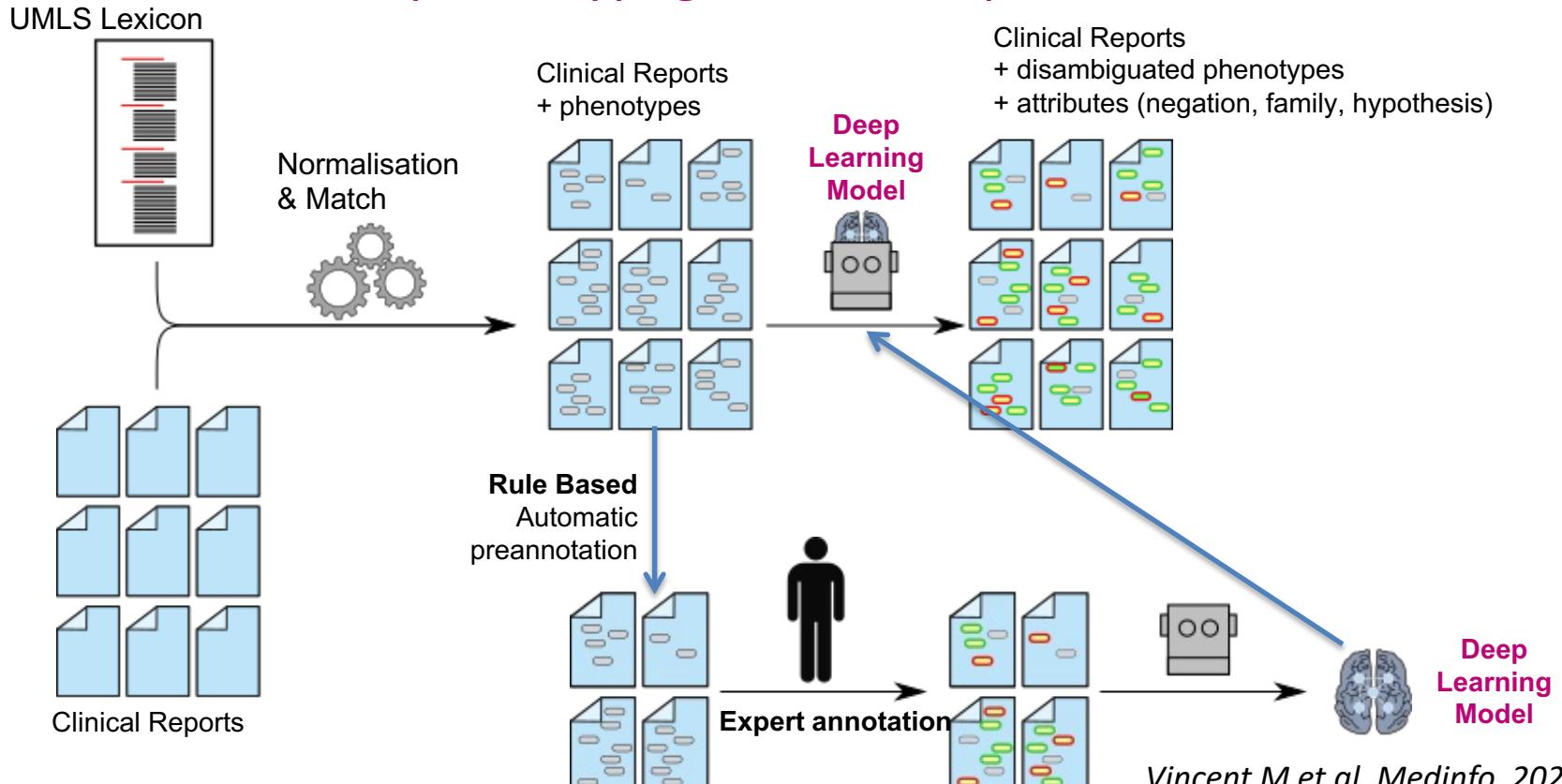
Extraire des informations depuis les séries temporelles (données labo)



Annotation d'un corpus par les méthodes à base de Règles

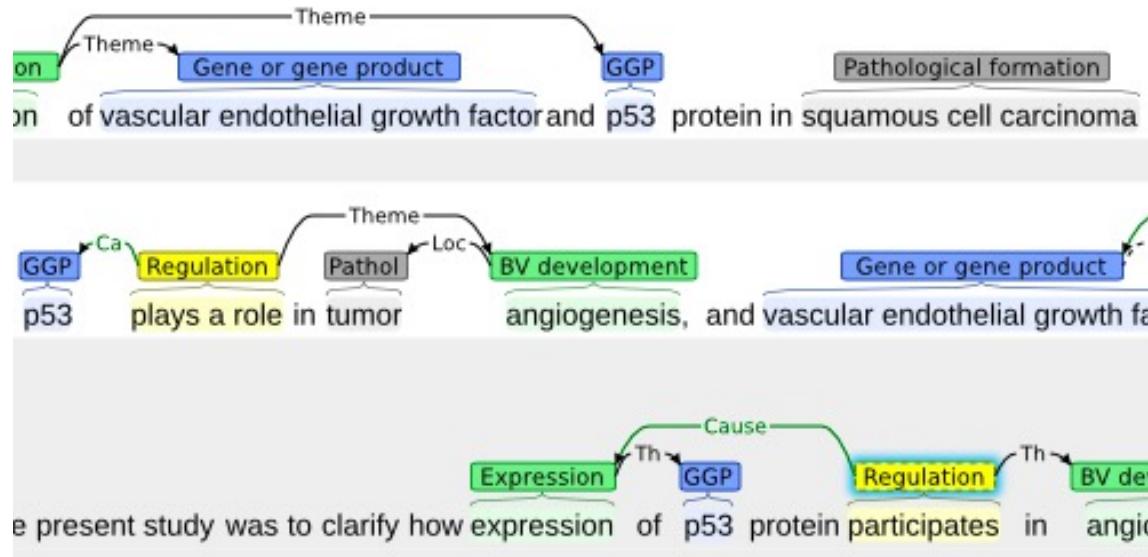


Utilisation de l'apprentissage profond pour le phénotypage automatique

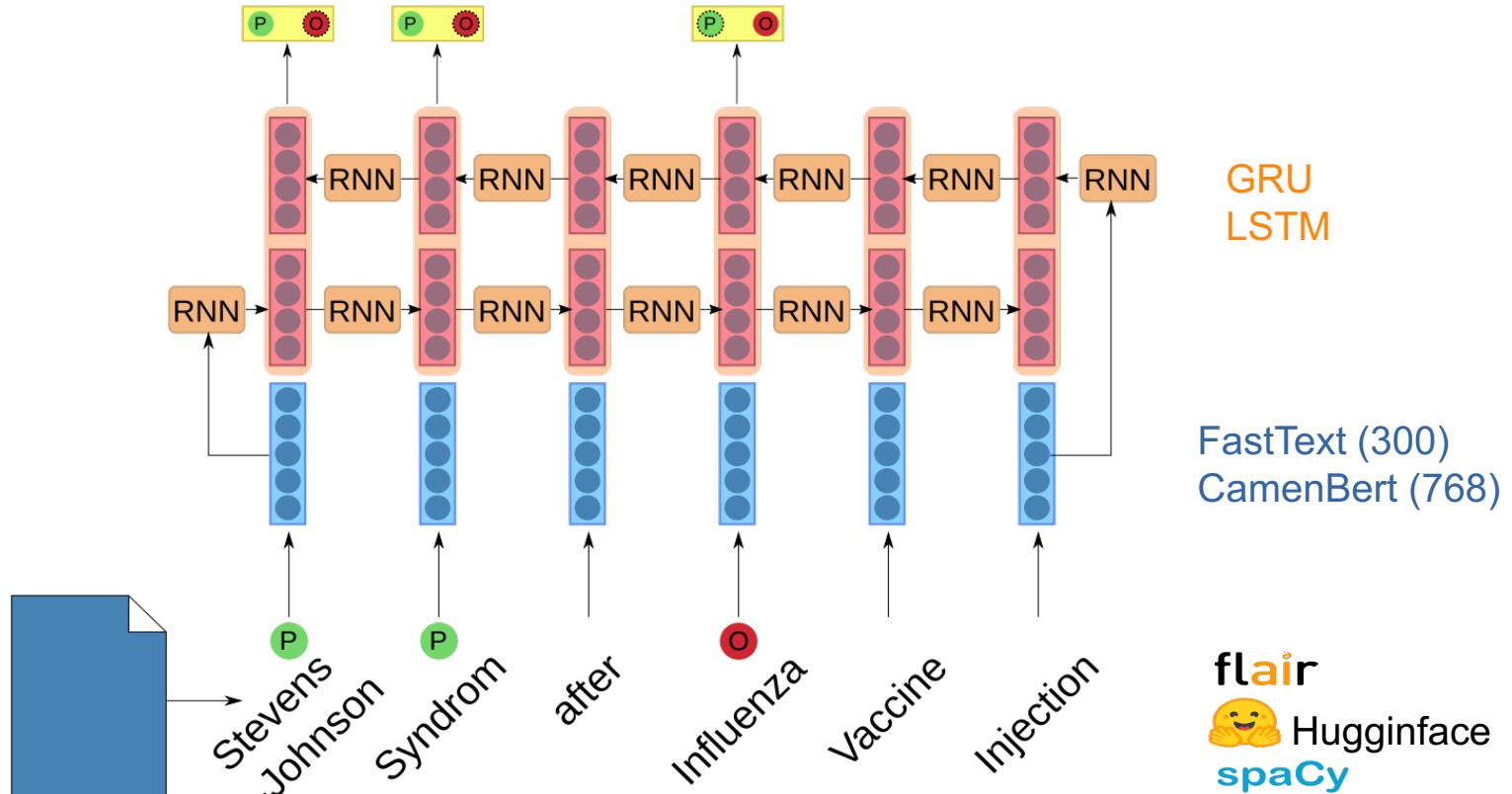


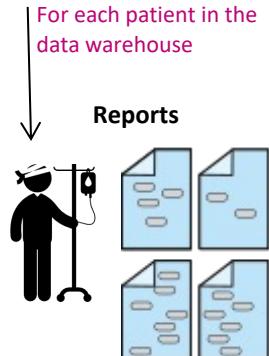
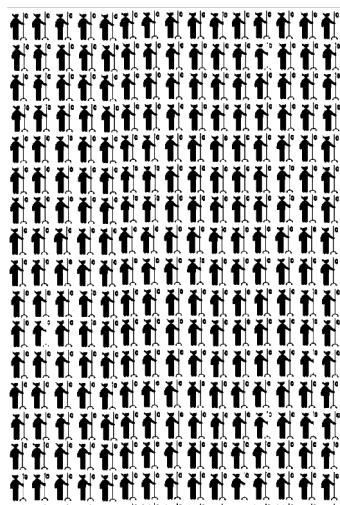
La partie pénible du traitement automatique du langage par deep learning

L'annotation manuelle des comptes rendus :



DrWarehouse+: Pipeline & Model

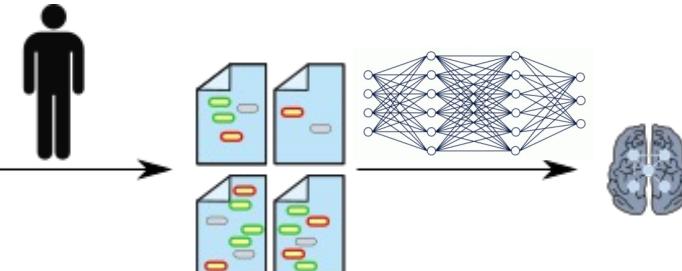




L'intelligence artificielle pour transformer le dossier patient en données structurées

Traitements du langage naturel pour détecter

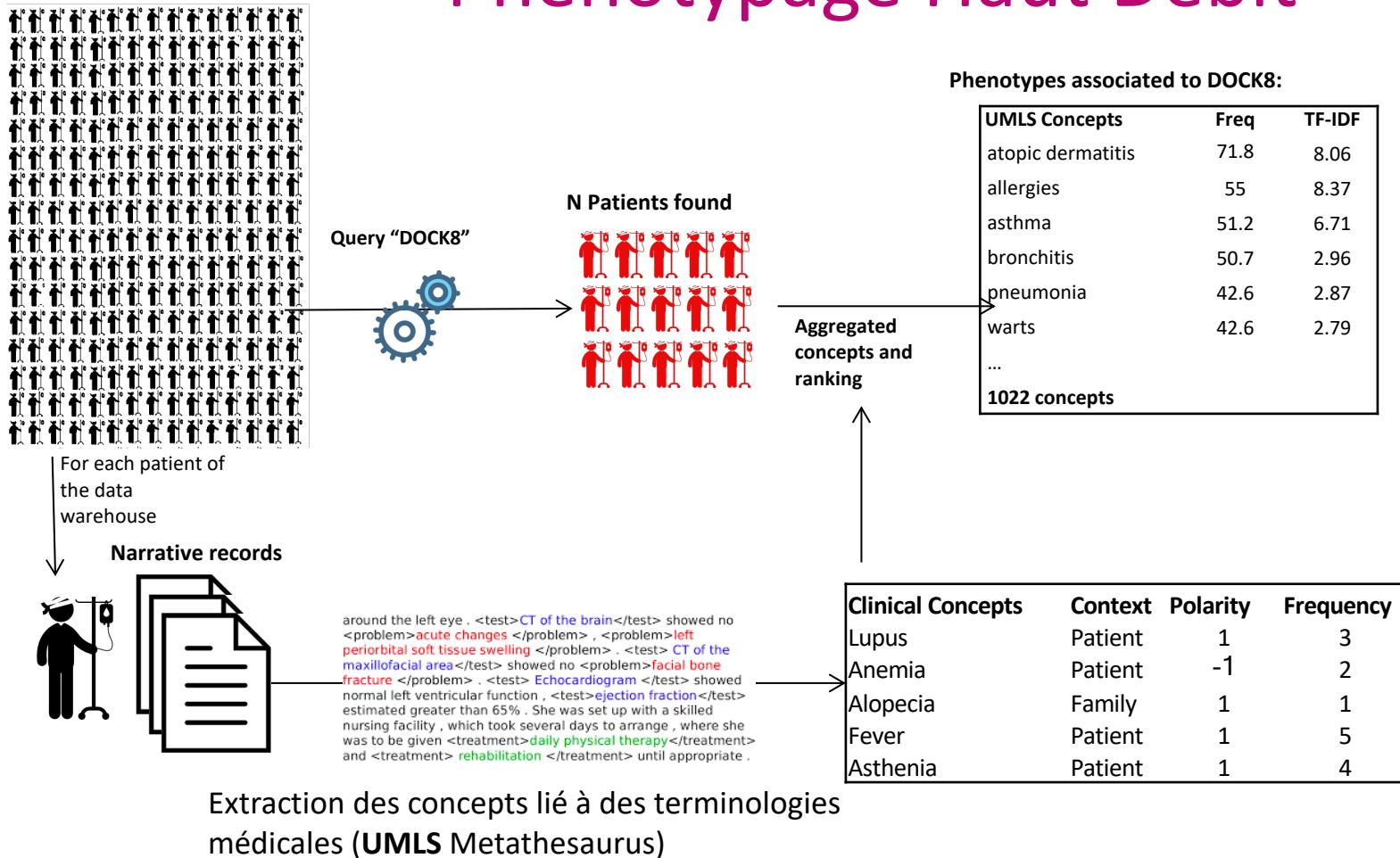
- Négation *Nous pouvons exclure le diagnostic de Syndrome de Rett*
- Antécédents familiaux *Le père est décédée d'un cancer du pancréas*
- Hypothèse *Suspicion de lupus. Recherche de mutation sur LRBA1*
- Extraire des phénotypes et des diagnostics
- Extraire informations depuis des photos (landmarks, morphométrie)
- Extraire des informations depuis les séries temporelles (données labo)



On améliore la précision de la détection de la négation de 85% à 95 % et les atcd familiaux de 41% à 87%

Data warehouse
830,000 patients

Phénotypage Haut Débit



Evaluation du phénotypage haut débit à base de règles

Garcelon et al. Orphanet Journal of Rare Diseases (2018) 13:85
https://doi.org/10.1186/s13023-018-0830-6

Orphanet Journal of
Rare Diseases

RESEARCH

Open Access



Next generation phenotyping using narrative reports in a rare disease clinical data warehouse

Nicolas Garcelon^{1,2,13*}, Antoine Neuraz^{2,3}, Rémi Salomon^{1,4}, Nadia Bahi-Buisson^{1,5}, Jeanne Amiel^{1,6,7}, Capucine Picard^{1,8,9}, Nizar Mahaoui^{1,8,10,11}, Vincent Benoit¹, Anita Burgun^{2,3,12} and Bastien Rance^{2,12}

Abstract

Background: 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.

Methods: 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.

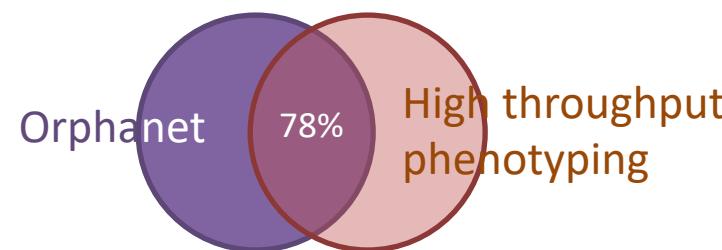
Results: Experts concluded that between 16 and 39 phenotypes could be considered as relevant in the top-50 phenotypes ranked by descending frequency discovered by Dr. Warehouse (resp. between 11 and 41 for TF-IDF). Average precision ranges from 0.55 to 0.91 for frequency and 0.52 to 0.95 for TF-IDF. 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.

Conclusions: Clinical Data Warehouses can be used to perform Next Generation Phenotyping, especially in the context of 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.

Keywords: Data warehouse, Next generation phenotyping, Data mining, Rare diseases, Natural language processing

Evaluation on 6 cohorts

- Bardet Biedl syndrome - 53 patients
- Activated PI3kinase delta syndrome (APDS) - 23 patients
- Rett syndrome - 209 patients
- DOCK8 deficiency - 15 patients
- Silver Russell Syndrome - 50 patients
- Lowe syndrome - 23 patients



Thesaurus francophones sont moins riches que thesaurus anglophone.

Nécessité de développer des approches d'enrichissement des thésaurus à partir des données réelles (comptes rendus)

Faviez, C. et al., 2022. Enriching UMLS-Based Phenotyping of Rare Diseases Using Deep-Learning: Evaluation on Jeune Syndrome. Challenges of Trustable AI and Added-Value on Health 844–848.

Rechercher des patients

Sur tout l'entrepôt

Syndrome de rett

280/280

Etendre aux synonymes :

+ Avancé - Réécrire la requête

+ Ajouter un filtre Full text

+ Ajouter un filtre structuré

+ Contraintes temporelles

+ Filtre patient

LANCER LA RECHERCHE

2

Sur tout l'hôpital :

195 Patients

473 Documents

Sur les documents trouvés - Sur tous les documents des patients trouvés

Profondeur : 10



Show 10 entries

| Concepts | # patients | FreqRes | PSS | Case-Weighted PSS |
|---------------------|------------|---------|------|-------------------|
| Syndrome de Rett | 122 | 62.6 | 55 | 1084.9 |
| marche apraxique | 37 | 19 | 48.7 | 254.05 |
| Hyperventilation | 54 | 27.7 | 13.8 | 130.61 |
| Bruxisme | 56 | 28.7 | 10.5 | 108.03 |
| Syndrome pyramidal | 88 | 45.1 | 4.9 | 97.59 |
| Osteoporose | 85 | 43.6 | 3.3 | 68.21 |
| Scoliose | 101 | 51.8 | 1.3 | 41.57 |
| Myoclonie | 43 | 22.1 | 3.1 | 28.83 |
| encephalopathie | 72 | 36.9 | 1.4 | 28.57 |
| Retard psychomoteur | 64 | 32.8 | 1.6 | 27.33 |

Phénotypage automatique

Extraction automatique de 50 millions de concepts médicaux à partir des comptes rendus hospitaliers de Necker. (Garcelon et al, OJRD, 2018), reliés à des ontologies (HPO, GO)

- Description automatique de cohortes
 - Génération d'hypothèse
 - Découverte de signes précoce pour le diagnostic de maladie rare (Lo Barco, OJRD, 2021)

⇒ Représentation automatique synthétique du dossier médical d'un patient en lien avec des ontologies

Data mining & extraction de données : Histoire naturelle

Syndrome de Myrhe

Extraction des 50 phénotypes les plus associés à Myrrhe avec DrWH

5 Associations phénotypiques jamais décrites dans OMIM et ORPHADATA



Yang, D.D., Baujat, G., Neuraz, A., Garcelon, N., Messiaen, C., Sandrin, A., Cheron, G., Burgun, A., Pejin, Z., Cormier-Daire, V., Angoulvant, F., 2020. Healthcare trajectory of children with rare bone disease attending pediatric emergency departments. *Orphanet J Rare Dis* 15, 2. <https://doi.org/10.1186/s13023-019-1284-1>

Search for patients

Across the entire data warehouse

Netherton X

Extend to synonyms :

+Advanced - Rewrite the query - Test the query - Pré calculer

- + Add a full text filter
- + Add a structured filter
- + Add a movement filter
- + Time constraints
- + Logical constraints
- + Patient filter

START A SEARCH

Query history

Show 20 entries Search:

Date Queries

- 27/11/2020 11:22 Filtered 1 : Documents containing 'syndrome de rett', Excluding negations
- 27/11/2020 09:05 Filtered 1 : Documents containing 'infection%' and 'eczema and thrombopenie' Extended to synonyms, Excluding negations
- 08/06/2022 00:35 Filtered 1 : Documents containing 'netherton', Excluding negations
- 07/06/2022 20:28 Filtered 1 : Documents containing 'netherton', Excluding negations
- 07/06/2022 08:15 Filtered 1 : Documents containing 'syndrome de rett', Excluding negations
- 01/06/2022 15:29 Filtered 1 : Documents containing 'syndrome de rett', Excluding negations
- 24/05/2022 11:13 Filtered 1 : Documents containing 'syndrome de rett', Excluding negations
- 24/05/2022 00:00 Filtered 1 : Documents containing 'syndrome de rett'

Result
Stats data
Phenotypes
DRG
Biology
Map
Clustering
Export data
Extraction

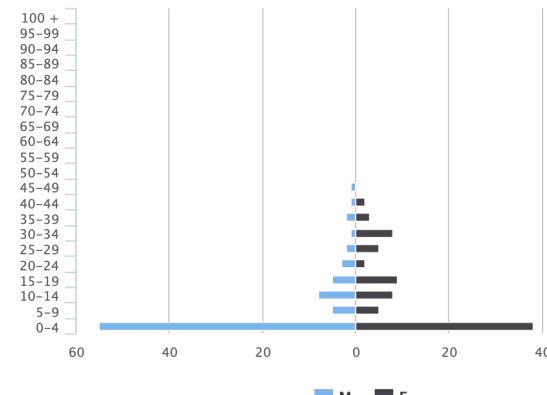
173 patients
1529 Documents
0 Movements

SAVE QUERY**SHARE YOUR QUERY**

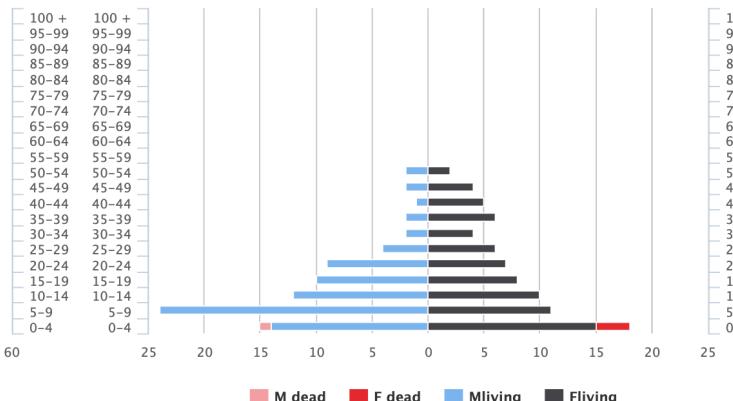
Description démographique automatique

- Statistics based on documents found

Population pyramid – age at the 1st document found



Population pyramid – today



Population pyramid – <20 years old in first document found



Population pyramid – <20 years old today



38 / 58

Rechercher des patients

Sur tout l'entrepôt

1

N

Etendre aux synonymes :

- + Ajouter un filtre Full text
 - + Ajouter un filtre structuré
 - + Ajouter un filtre mouvement
 - + Contraintes temporelles
 - + Contraintes logiques
 - + Filtre patient

[Historique requête](#)

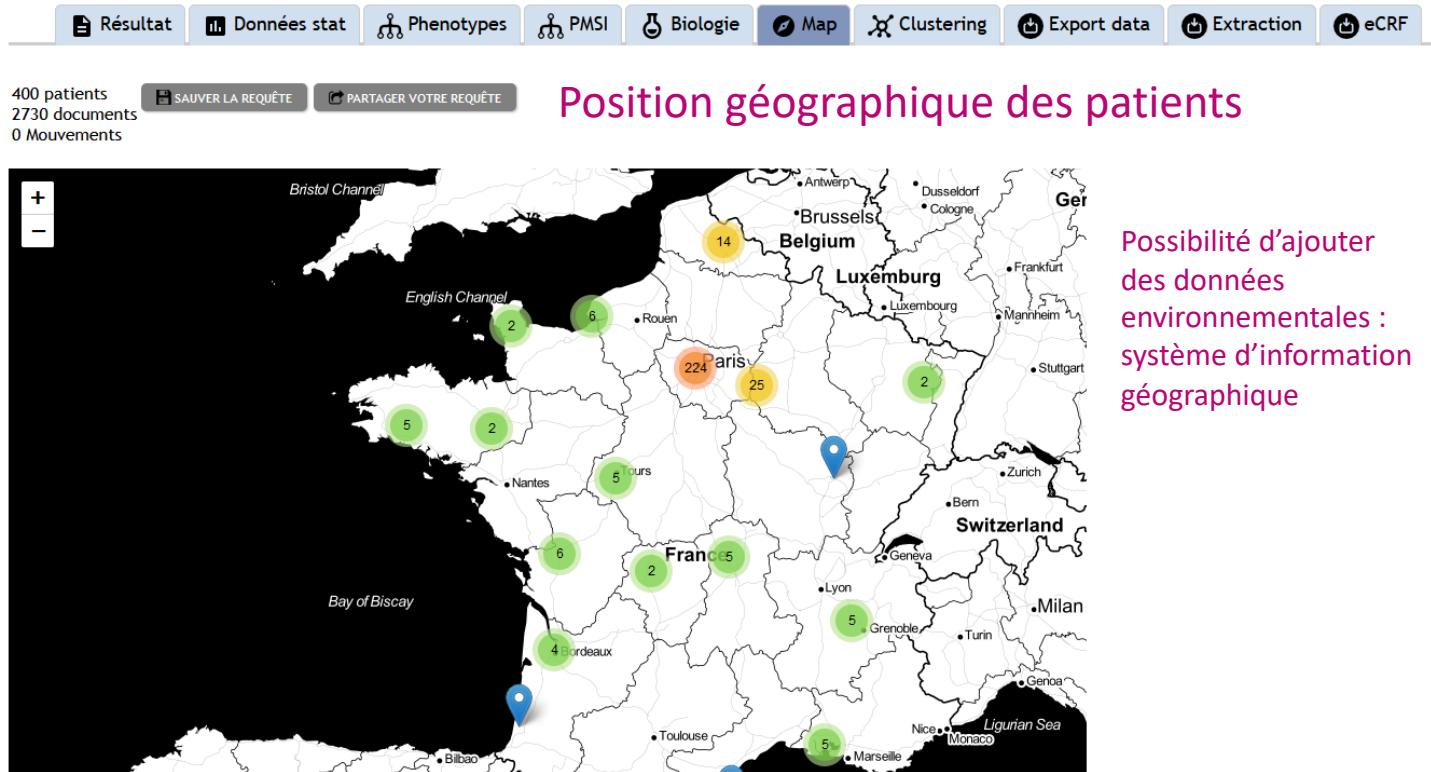
Show **20** entries Search: Requêtes

27/11/2020 Filtre 1 : Documents contenant 'syndrome de rett' , en excluant les négations
11:22

27/11/2020 Filtre 1 : Documents contenant 'infection%' and 'eczema' and 'thrombopenie' Etendu aux synonymes , en excluant les négations
09:05

29/11/2021 Filtre 1 : Documents contenant 'osteosarcome' , en excluant les négations
15:15

29/11/2021 Filtre 1 : Documents contenant 'cancer du pancréas' , âgés de 0 ans à 20 ans la date du document , en excluant les négations
14:18



Possibilité d'ajouter des données environnementales : système d'information géographique

Analyse du parcours de soin d'une population

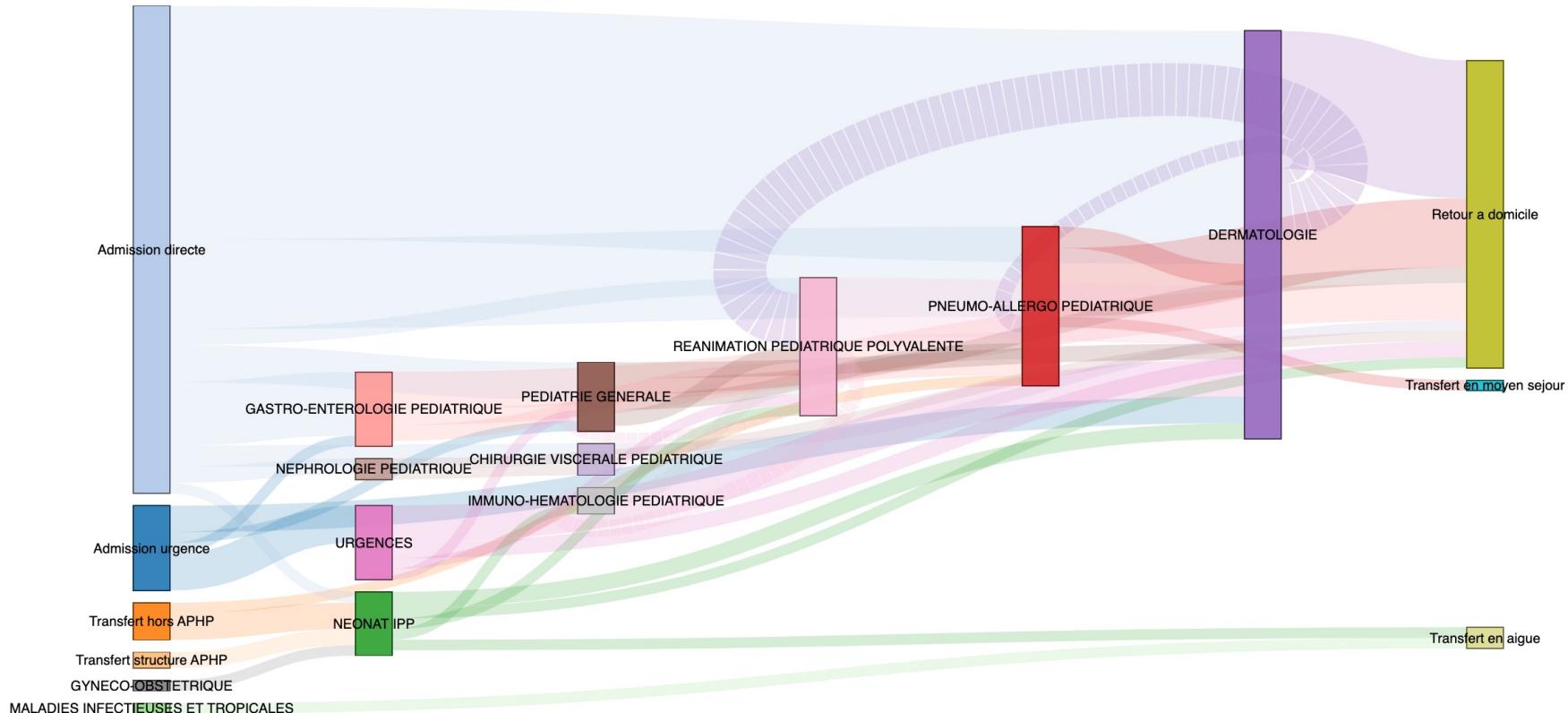


Illustration des mouvements d'une population de patients au sein d'un hôpital

40 / 58

Analyse du parcours de soin d'une population

Nouvelles recommandations en analysant les mouvements des patients atteintes d'une maladie osseuse rare t-7 jours avant et t+30 jours après une visite aux urgences pédiatriques de l'hôpital Necker.

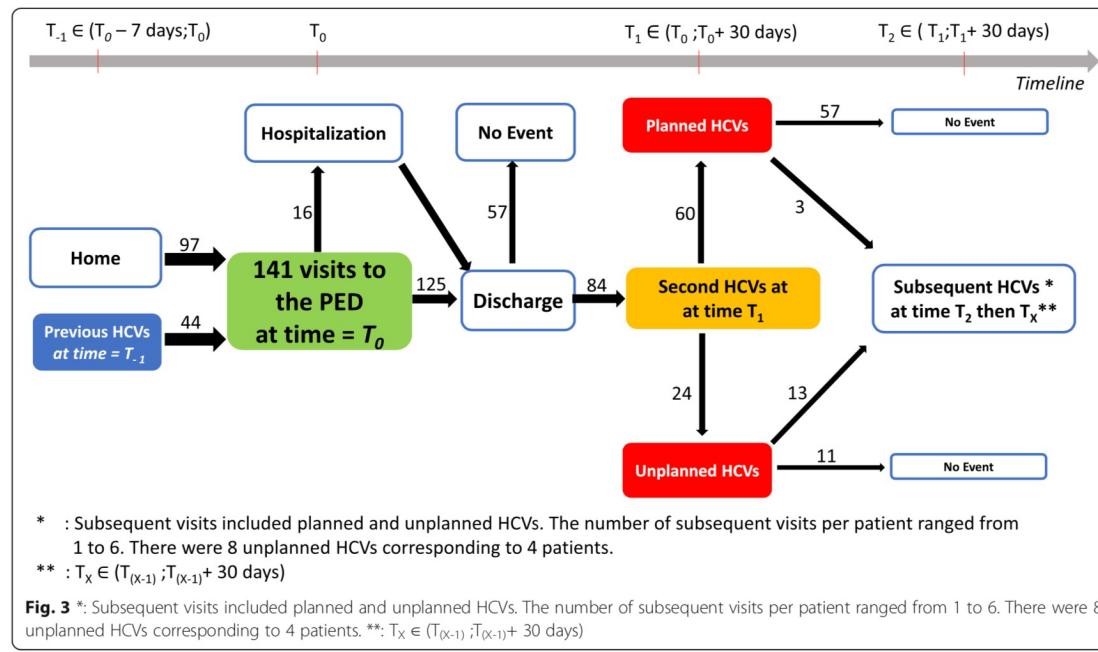
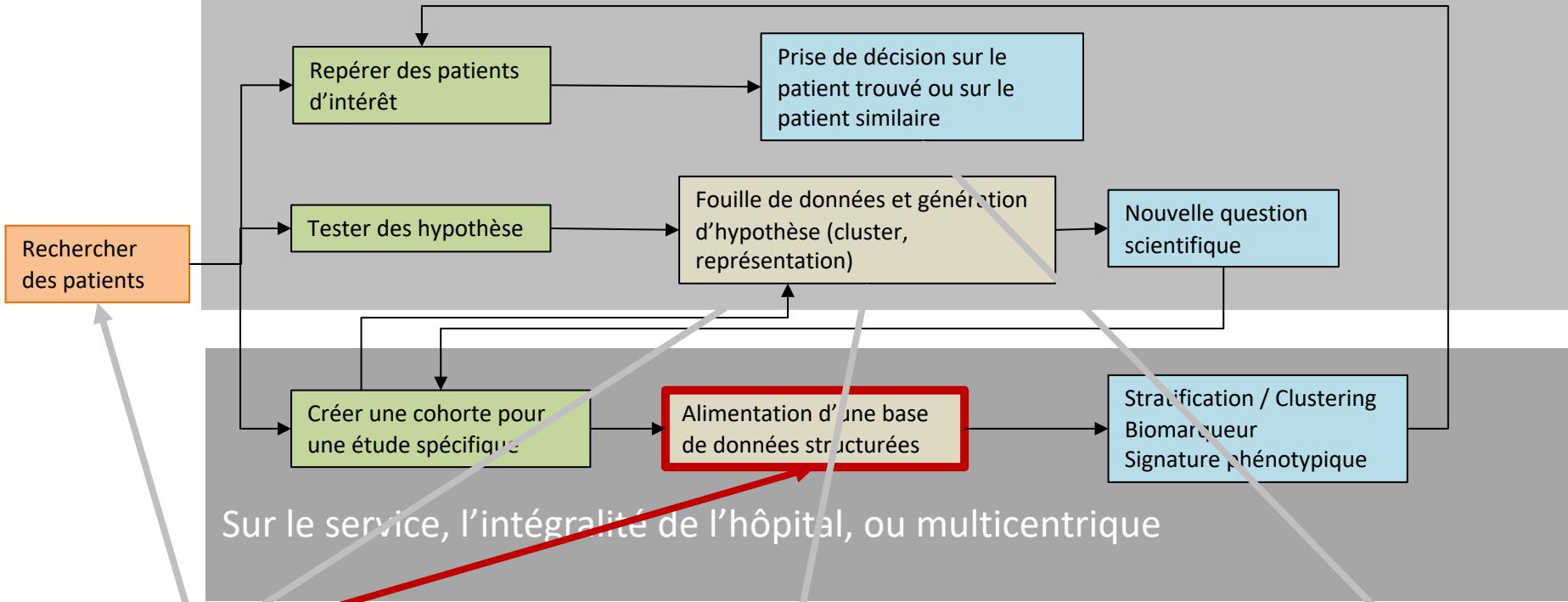


Fig. 3 *: Subsequent visits included planned and unplanned HCVs. The number of subsequent visits per patient ranged from 1 to 6. There were 8 unplanned HCVs corresponding to 4 patients. **: $T_x \in (T_{(X-1)}; T_{(X-1)} + 30 \text{ days})$

Dans le service du médecin



Traitements automatiques
du langage naturel /
Images / signal

Ajout de connaissances externes (thèses, ontologie, etc) : enrichissement et normalisation.
Données épidémiologiques

Rendre les données des patients compatibles aux outils d'aide à la décision

Une mise en qualité des données selon les besoins

Accélérer la recherche, rationnaliser et automatiser le processus de collecte.

Saisie semi automatique dans les eCRF : **Le Smart Data Extractor**

- ① Paramétrage d'un eCRF dans Dr WH: définition de la restructuration des données
- ② Extraction sur le dossier d'un patient en conservant le lien avec les comptes rendus : contextualisation
- ③ Contrôle qualité, correction et validation
- ④ Export sécurisé et anonymisé

Le Smart Data Extractor : Paramétrage des items dans l'entrepôt

Dr WareHouse ©Imagine
Entrepôt de données

Accueil | Moteur de recherche | Mes requêtes | Mes Cohortes | Outils | Mes ecrf | Patient Nom / IPP | Admin | Notifications 5 | Nicolas Demo Garcelon | deconnexion

Mes Ecrf

+ Créer un Ecrf Ecrf : SNIF

Mes Ecrf :

Search:

Description Les items patients

+ Ajouter les items par bloc

APDS 0

ERKNET et NDI 22

rett 7

SNIF 23

Période d'extraction des données du au
Intervalle d'âge (en années) pour l'extraction des données de à

| Ordre | Question | Type | Valeur | Pattern | Pattern Index | Fonctions existantes | Source de document | Local Codes Rechercher un code | Noms Externes | Codes Externes | Période | X |
|-------|------------------------------------|---------|---|---|---------------|----------------------|--------------------|------------------------------------|---------------|----------------|---------|---|
| 1 | Date de corticothérapie | date | | corticotherapies? cortancyl prednisol steroides? corticoïdes? | | | | | | | first | |
| 2 | Age à la corticothérapie | numeric | | corticotherapies? cortancyl prednisol steroides? corticoïdes? | | | | | | | first | |
| 3 | Nb jours depuis la dernière visite | numeric | | | | | | Nb jours depuis la dernière visite | | | all | |
| 4 | Perdu de vu | radio | oui non | | | | | | • | • | all | |
| 7 | Poids au diagnostic | numeric | corticotherapies? cortancyl prednisol steroides? corticoïdes? | (poids pese poid poids a poids de poids etait de poids est de)[^a-z0-9]*([0-9]+[.,]?[0-9]*) *kg | 2 | | | | | | first | |
| 8 | Taille au diagnostic | numeric | | (mesure taille taille de taille a)[^a-z0-9]*([0-9]+[.,]?[0-9]*) *cm | 2 | | | | | | first | |
| 9 | BMI au diagnostic | numeric | | BMI[^a-z0-9]*([0-9]+[.,]?[0-9]*] | 1 | | | | | | first | |
| 10 | Particularités cliniques | list | Diabète Pathologies sous-jacentes Syndromique | Diabète Diabète diabète diabète Pathologies sous-jacentes Pathologies sous-jacentes pathologie sous-jacentes Pathologies sous-jacente Pathologie sous-jacente syndromique syndrome +de microcéphalie retard de dysmorphie | | | | | • | • | all | |
| 11 | Facteur déclenchant | list | allergie infection | allergie allergies asthme eczema eczéma rhinite allergique atopie infection infections | | | | | • | • | all | |

44 / 58

Le Smart Data Extractor

Accélérer la recherche, automatiser le processus de collecte sur le dossier d'un patient :

Automatiser le processus de collecte

- Accélération par de l'aide à la saisie : repérage des comptes rendus contenant l'information
- Contrôle qualité et correction
- Sauvegarde
- Interopérabilité avec REDCap : transfert de données

Evaluation sur 80 patients, 20 variables

- Un TEC, un cardiologue
- **Divise par 2 le temps de saisie**
- **Exhaustivité 64% -> 71%**
- **Divise par 4 le nombre d'erreurs**

Quennelle S et al. The Smart Data Extractor, a Clinician Friendly Solution to Accelerate and Improve the Data Collection During Clinical Trials. Stud Health Technol Inform. 2023 May 18;302:247-251. doi: 10.3233/SHTI230112. PMID: 37203656.

Patient M,
9 ans

documents Biologie TimeLine Parcours PMSI Cohorte Concepts Similarité eCRF

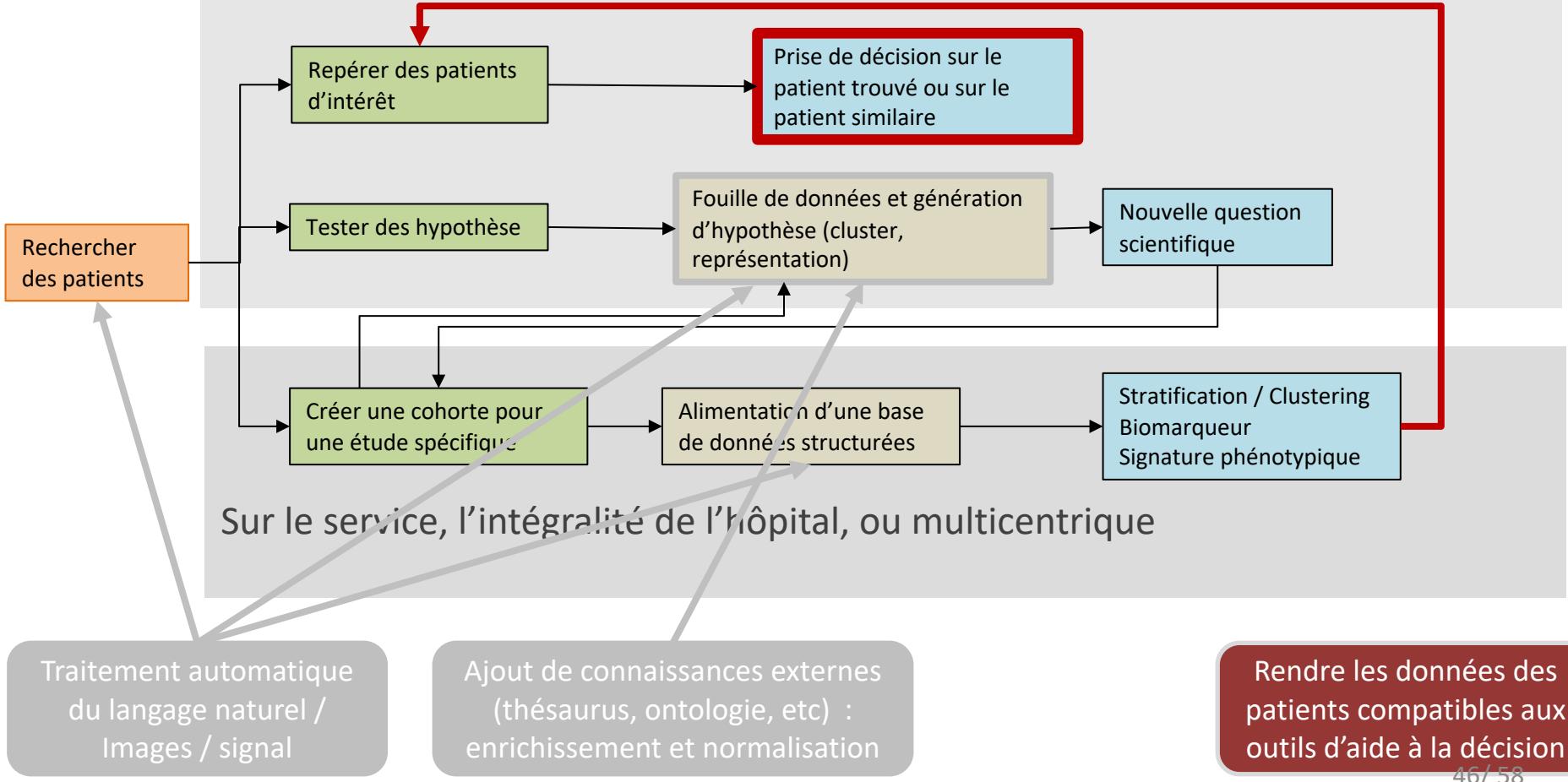
ERKNET et NDI x ▾

*Vous pouvez réaliser plusieurs extractions pour différents suivis du patient.
Vous devez préciser pour chaque suivi une date autour de laquelle les données doivent être extraites,
ainsi qu'un nombre de jours maximum avant et après cette date
+ ajouter un suivi*

DISPLAY EXTRACT

| | | |
|---|--|---|
| 3. Please enter a patient's pseudonym | <input type="text"/> | ✓ |
| 4. Current age (years) | : 9 <input type="text" value="9"/> (auto) | ✓ |
| 5. Age at diagnosis test | <input type="text"/> | ✓ |
| 6. Please select gender | <input type="radio"/> Female <input checked="" type="radio"/> Male (auto) | ✓ |
| 7. Poids actuel | [DATE] : 14,3 <input type="text" value="14,3"/> (auto) | ✓ [DATE] iniquement Son poids est de 14,3 kg pour une |
| 8. Taille actuelle | [DATE] : 98,7 <input type="text" value="98,7"/> (auto) | ✓ [DATE] pour une taille de 98,7 cm La cr |
| 9. Ethnicity? (Please select all that apply.) | <input type="checkbox"/> American Indian or Alaskan Native <input type="checkbox"/> Asian or Pacific Islander <input type="checkbox"/> Black or African America <input type="checkbox"/> Hispanic or Latino <input type="checkbox"/> White or Caucasian <input type="checkbox"/> Prefer not to answer | ✓ |
| 10. Genetics | <input checked="" type="checkbox"/> AVPR2 (auto) <input type="checkbox"/> AQP2 <input type="checkbox"/> Tested but no mutation <input type="checkbox"/> Unknown | ✓ [DATE] nique avec mutation AVPR2 identifiee |
| 11. Complication de l'arbre urinaire | <input checked="" type="checkbox"/> Hydronephrose (auto) <input checked="" type="checkbox"/> Dysfonctionnement de la vessie (auto) <input type="checkbox"/> Autre | ✓ [DATE] y a pasde dilatation des cavit [DATE] Pas de dilatation des cavit [DATE] Pas de dilatation des cavit [DATE] nale et vesicale : reins e [DATE] nale et vesicale : reins e |

Dans le service du médecin



Des fonctionnalités centrées patient

Faciliter l'exploration du dossier du patient

The screenshot shows the Dr WareHouse software interface. At the top, there is a header with the title "Dr WareHouse ©Imagine" and a sub-header "Entrepôt de données". Below the header is a navigation bar with links: Accueil, Moteur de recherche, Mes requêtes, Mes cohortes, Mes demandes, Outils, Patient Nom / IPP, Notifications, and Nicolas Garcelon An.

The main content area displays a search result for "Patient M," with the search term "biopsie" entered in a search bar and a "RECHERCHER" button.

The results section shows 38 documents found, with a preview of the first document:

ORBIS CR - courrier
Histologie de la **biopsie** réalise le [REDACTED] : lesions de [...]

DIAMIC Biopsie cutanée avec colo spé pour diag d'affection non carcinologique

La **biopsie** communiquée fixée a été analysée sur plusieurs niveaux de coupe [...] Le reste de la **biopsie** est sans particularité [...]

ORBIS CR - courrier
Biopsie cutanée réalisée le [REDACTED] : aspect histologique observe montre d'une part des lesions de [...] Réalisation de **biopsies** cutanées au bloc opératoire sous anesthésie générale le [...] Bilan dermatologique en attente (**biopsies** cutanées en cours) [...]

ORBIS CR - courrier
Dernière **biopsie** cutanée en sept [...] **Biopsies** : aspect de vascularite leucocytoclasique malgré l'absence de nécrose fibrinoïde [...] Prevoir une **biopsie** cutanée pour nouvelle histologie + IFD plus ou moins microscopie électronique si [...] réalisation dans le même temps de la FOGD une **biopsie** cutanée des lesions de purpura des [...]

ORBIS CR - courrier
Biopsie cutanée en 09/2014 (IFD sur lesion ancienne --> difficilement interprétable): dépôts IgG et C3 [...] **Biopsie** : aspect de vascularité leucocytoclasique malgré l'absence de nécrose fibrinoïde [...] AG prevue en [...] pour une endoscopie : prévoir **biopsie** cutanée pour nouvelle histo + IFD [...]

ORBIS CR - courrier
Biopsie cutanée en [...] (IFD sur lesion ancienne --> difficilement interprétable): dépôts IgG et C3 [...] **Biopsie** : aspect de vascularité leucocytoclasique malgré l'absence de nécrose fibrinoïde [...] AG prevue en [...] pour une endoscopie : prévoir **biopsie** cutanée pour nouvelle histo + IFD [...]

ORBIS CR - courrier
- **Biopsie** cutanée : aspect histologique, malgré l'absence de nécrose fibrinoïde, fait conclure à une vascularite [...] Il n'y a pas eu de **biopsie** [...] Durant le même geste anesthésique, une PBH et une **biopsie** cutanée ont été réalisées [...]

On the right side of the screen, there is contact information for the laboratory:

SERVICE D'ANATOMIE ET DE CYTOLOGIE PATHOLOGIQUES
Groupe Hospitalier Necker-Enfants Malades
149 rue de Sevres 75743 PARIS CEDEX 15

Secretariat : [TEL]
Fax : [TEL]

Recu le : [DATE]
GASTRO-ENTEROLOGIE ENF
Patient : [NOM]
[PRÉNOM]
Sexe : M
NIP : [IPP] / NDA : [REDACTED]

A la demande du : [REDACTED]
Copie au(x) :
NSLE / NCGE

NECKER-ENFANTS MALADES
75743 PARIS CEDEX 15 age de [DATE]

RENSEIGNEMENTS CLINIQUES :
Pousse de purpura non infiltrée des membres. Contexte de maladie de Hirschsprung avec thrombopenie fluctuante. Purpura rhumatoïde ? Anomalie du tissu élastique ?

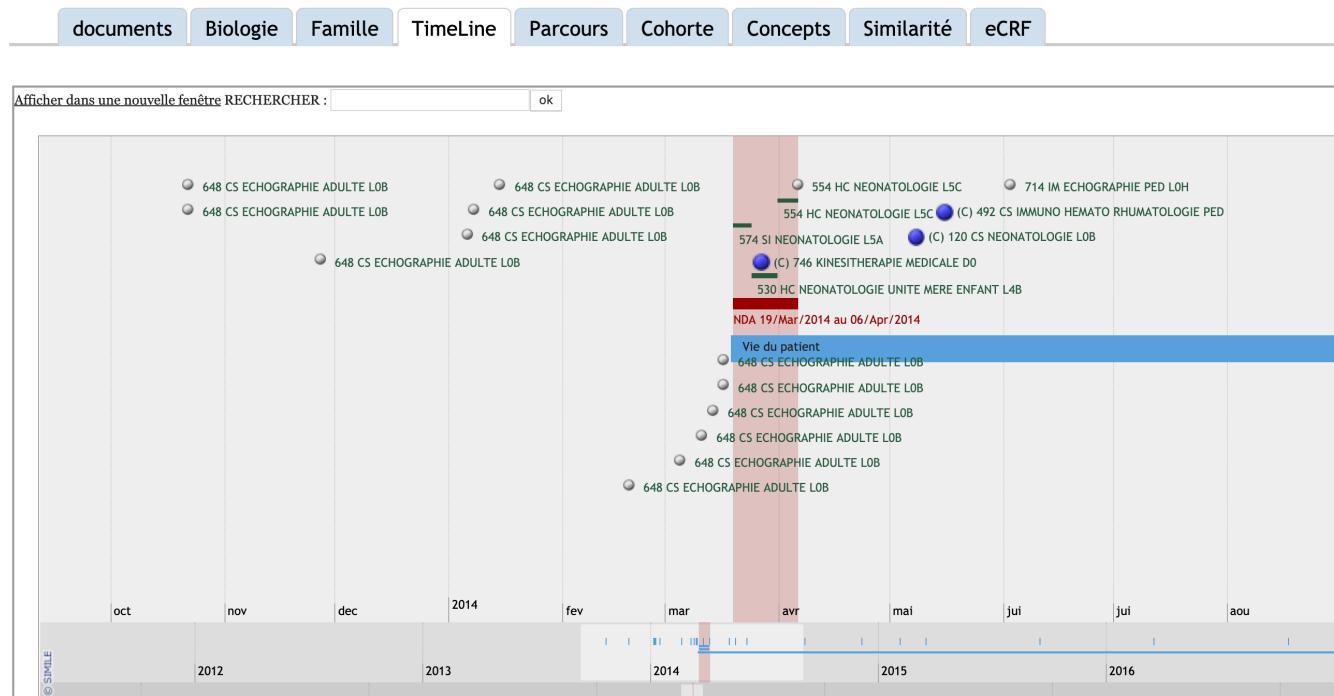
NATURE DU PRELEVEMENT :

Des fonctionnalités centrées patient

Visualisation du parcours du patient dans **le temps**

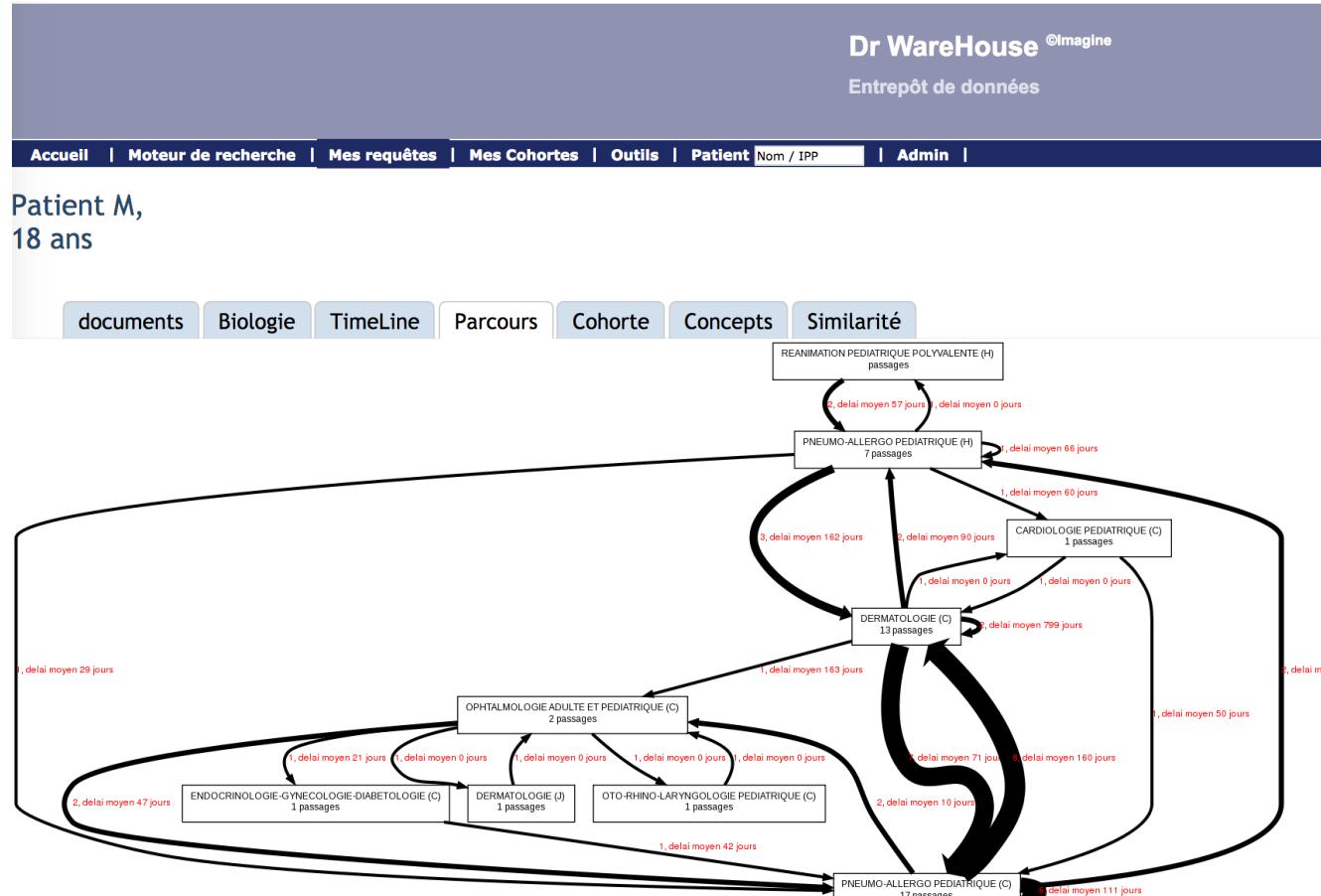
Patient F, 5 ans

Direct access to the next patient of the cohort Test similarite



Des fonctionnalités centrées patient

Visualiser le parcours de soin d'un patient dans l'espace



Recherche translationnelle et similarité

Patients non diagnostiqués



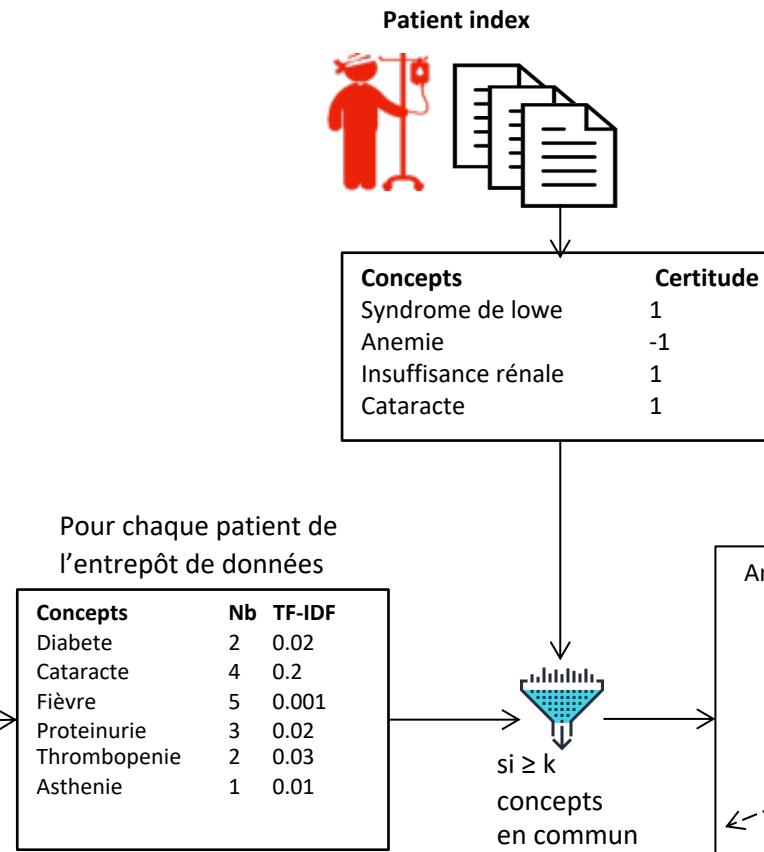
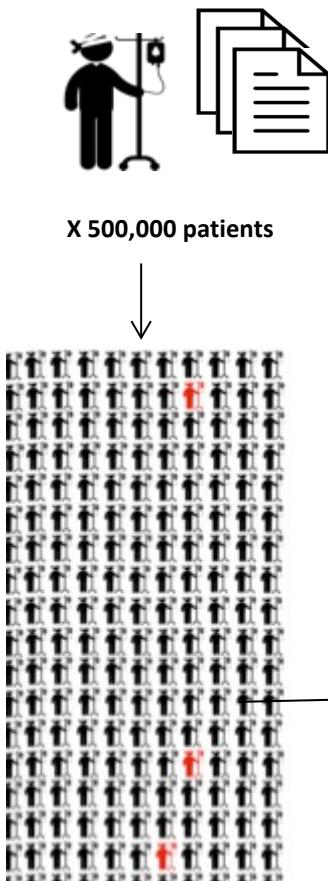
Un patient diagnostiqué
Phénotype complexe



Similarité avec le
patient index
Proximité phénotypique

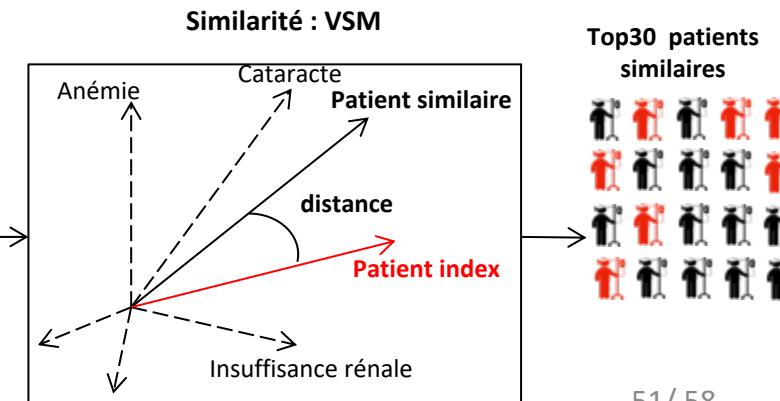


Recherche translationnelle et similarité



Dimensions définies par les concepts du patient index

$$\cos(\vec{P}_{index}, \vec{P}_k) = \frac{\vec{P}_{index} \cdot \vec{P}_k}{\|\vec{P}_{index}\| \|\vec{P}_k\|}$$



Développement d'algorithmes innovants pour accélérer l'aide au diagnostic

Similarité phénotypique :

Développer une métrique permettant de calculer une distance phénotypique entre les patients à partir de leurs comptes rendus hospitaliers (2017)

Prise en compte de l'ontologie HPO pour améliorer le calcul (2022) en utilisant la hiérarchie

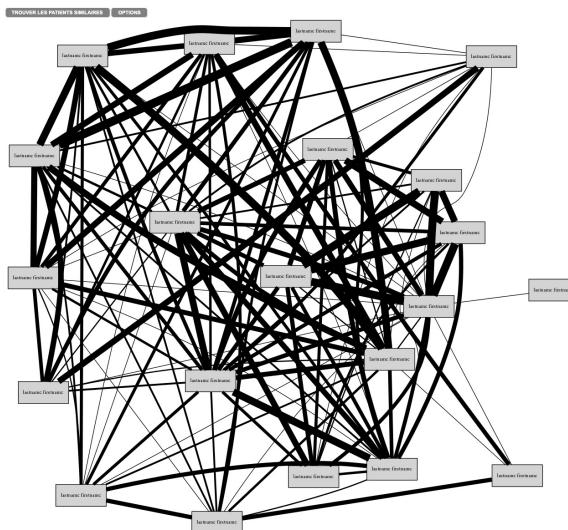
Pour aller au final sur des méthodes d'embedding

Garcelon N., Neuraz A., Benoit V., Salomon R., Kracker S., Suarez F., Bahi-Buisson N., Hadj-Rabia S., Fischer A., Munnich A., Burgun A., **2017**. 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, 51–61

Chen X., Faviez C., Vincent M., Briseño-Roa L., Faour H., Annereau J.-P., Lyonnet S., Zaidan M., Saunier S., Garcelon N., Burgun A., **2022**. Patient-Patient Similarity-Based Screening of a Clinical Data Warehouse to Support Ciliopathy Diagnosis. *Frontiers in Pharmacology* 13.

Nom Prénom né le 01/01/2000

Documents Biologie TimeLine Parcours Cohorte Concepts Similarité



› Mise en production dans l'entrepôt de Necker :
Similarité entre patients basée sur les comptes
rendus des patients



BRIEF COMMUNICATION

Deep phenotyping unstructured data mining in an extensive pediatric database to unravel a common *KCNA2* variant in neurodevelopmental syndromes

Marie Hully¹, Tommaso Lo Barco¹, Anna Kaminska^{1,2}, Giulia Barcia^{1,3}, Claude Cances⁴, Cyril Mignot⁵, Isabelle Desguerre¹, Nicolas Garcelon^{6,7}, Edor Kabashi⁸ and Rima Nabbout^{1,8,✉}

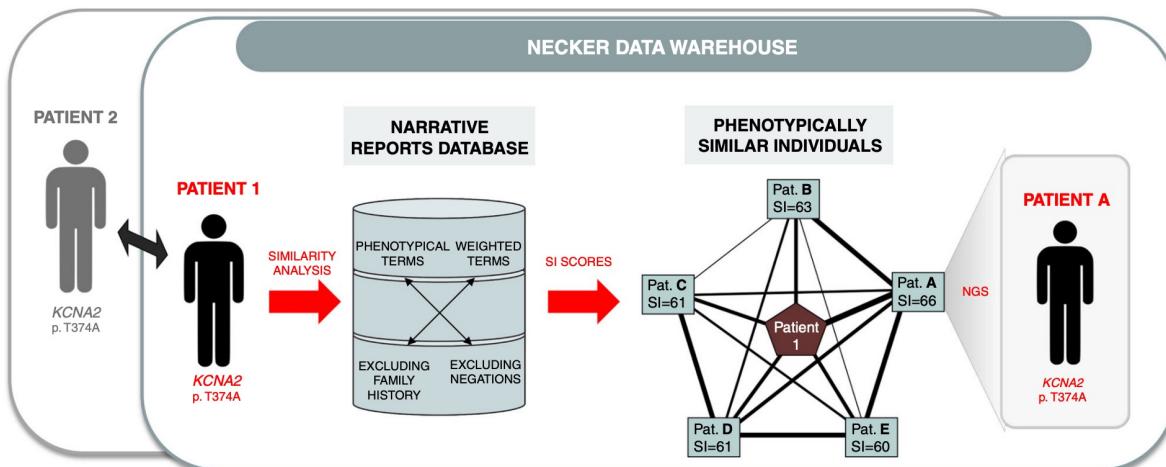


Fig. 1 Display of the two patients (patient 1 from our institution and patient 2 from another institution in our reference center network) sharing the same phenotype and the same *KCNA2* variant. Similarity analysis with all data warehouse narrative reports was performed, yielding a high similarity index (SI) in five patients (patients A-E). Exome sequencing validated that patient A, who had the highest SI, harbored the same *KCNA2* variant. NGS next-generation sequencing.

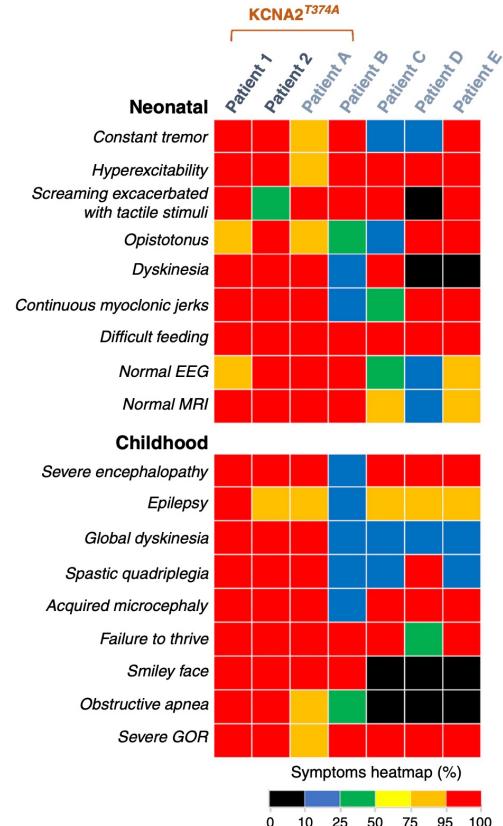


Fig. 2 Clinical heat map describing the detailed characteristics of the patients in this study. Heatmap for patient 1 and 2 with

Dr Warehouse : un entrepôt hospitalier open source

Journal of Biomedical Informatics 80 (2018) 52–63

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

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

Nicolas Garcelon^{a,b,*}, Antoine Neuraz^{b,c}, Rémi Salomon^{a,d}, Hassan Faour^a, Vincent Benoit^a, Arthur Delapalme^a, Arnold Munnoch^{a,e,f}, Anita Burgun^{b,c}, Bastien Rance^{b,g}

^a Institut Imagine, Paris Descartes Université Paris Descartes-Sorbonne Paris Cité, Paris, France

^b INSERM, Centre de Recherche des Cordeliers, UMR 1138 Equipe 22, Université Paris Descartes, Sorbonne Paris Cité, Paris, France

^c Département of Medical informatics, Hôpital Necker-Enfants Malades, Assistance Publique des Hôpitaux de Paris, Paris, France

^d Service de Néphrologie Pédiatrique, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France

^e Département de génétique médicale, Hôpital Necker-Enfants Malades, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France

^f Centre de Référence des Maladies Osseuses Constitutives, 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

^g Hôpital Européen Georges Pompidou, Assistance Publique-Hôpitaux de Paris (AP-HP), Université Paris Descartes, Sorbonne Paris Cité, France

ARTICLE INFO

ABSTRACT

Keywords:
Software
Computational biology
Method
Data warehouse
Rare diseases
Electronic health records
Information storage and retrieval
Text-mining

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 2837 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/

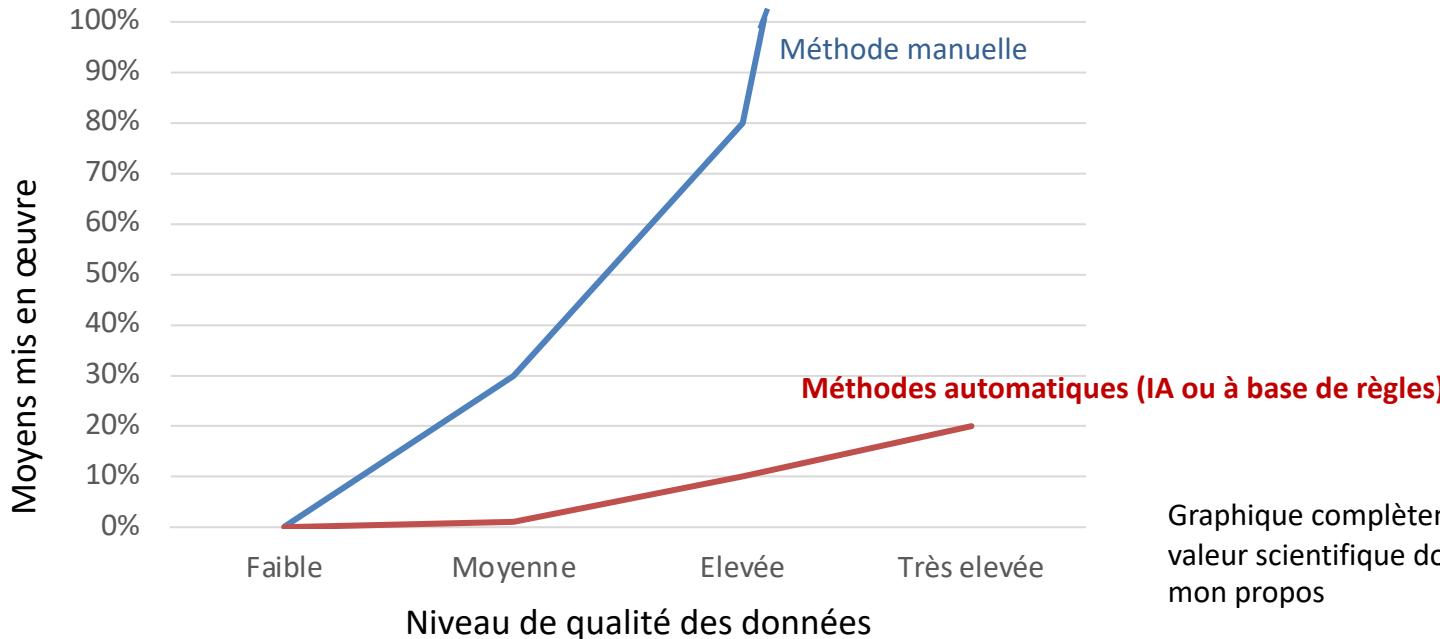
- Réappropriation des données cliniques par les producteurs de données (cliniciens)
- Accélérer l'accès à la mémoire collective dans le quotidien du médecin
- Accélérer la réutilisation des données pour la recherche

Création de codoc, spinoff d'imagine en 2017
Dr Warehouse est dans 12 hôpitaux en France

Deuxième verrou à la réutilisation : la mise en qualité des données

Le niveau de qualité dépend de l'objectif et des moyens à mettre en œuvre pour y arriver.

L'émergence des grands modèles de langage (LLM) permettrait de restructurer à moindre coût (excepté le coût d'apprentissage + GPU)

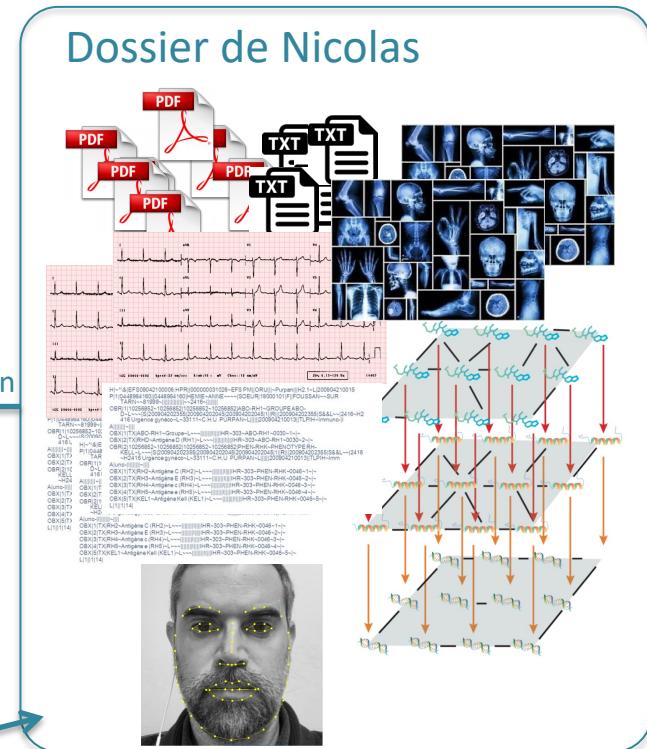
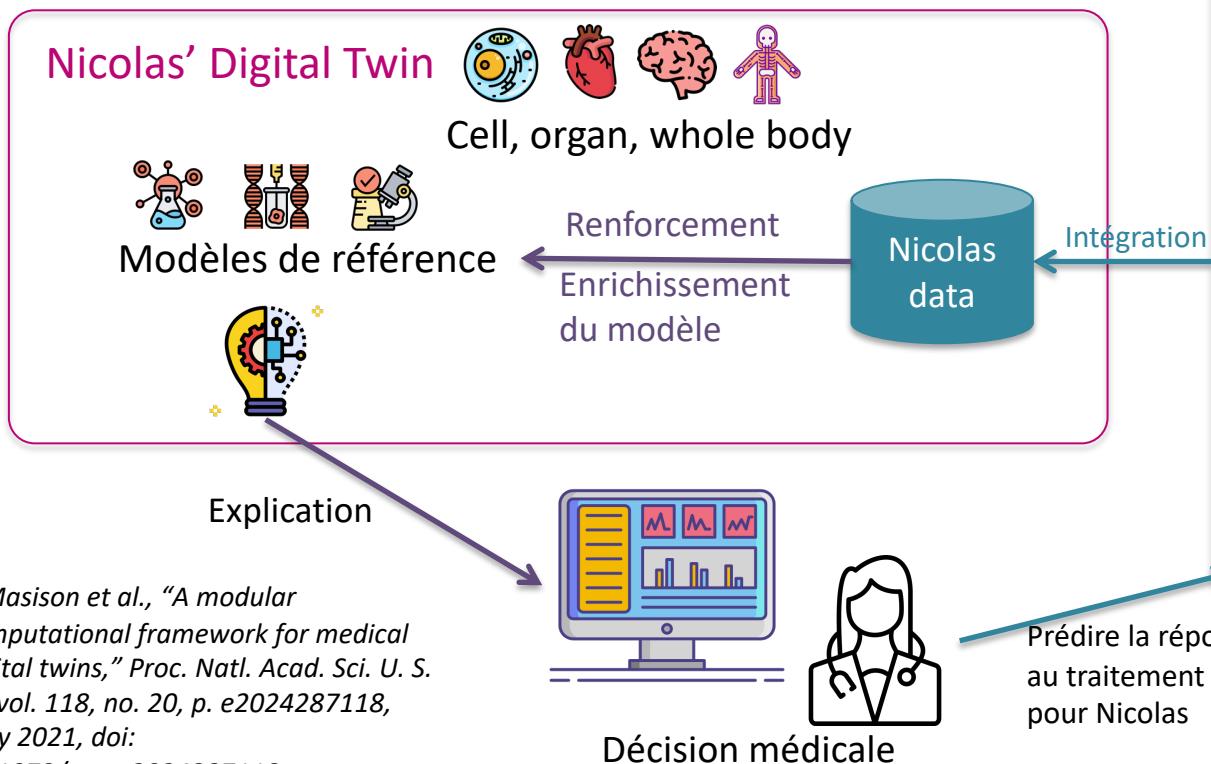


Graphique complètement inventé sans aucune valeur scientifique dont l'objectif est d'illustrer mon propos

Le système de santé apprenant future : le jumeau numérique

Tester des décisions : “Et si ...”

3 composants : Données + Modèle + Simulation



Conclusion en 4 points

1. Développer des entrepôts de **données brutes** : pour garder le contexte et le sens
2. Développer des outils de fouilles de données pour faciliter l'exploration des données.
3. Développer des outils d'extraction pour répondre à une question scientifique précise.
4. Développer des interfaces pour exécuter les algorithmes d'aide sur le dossier des patients.

Avoir une approche pragmatique et réaliste de réutilisation des données de vie réelle

Nicolas.garcelon@institutimagine.org