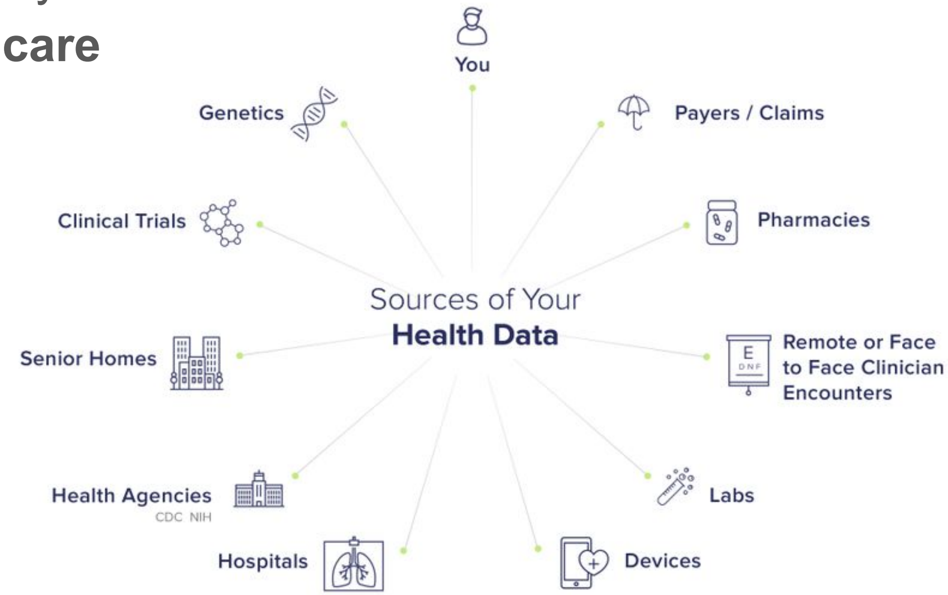
Data secondary use

Master Public Health Data Science

Vianney Jouhet, MD, PhD
vianney.jouhet@u-bordeaux.fr

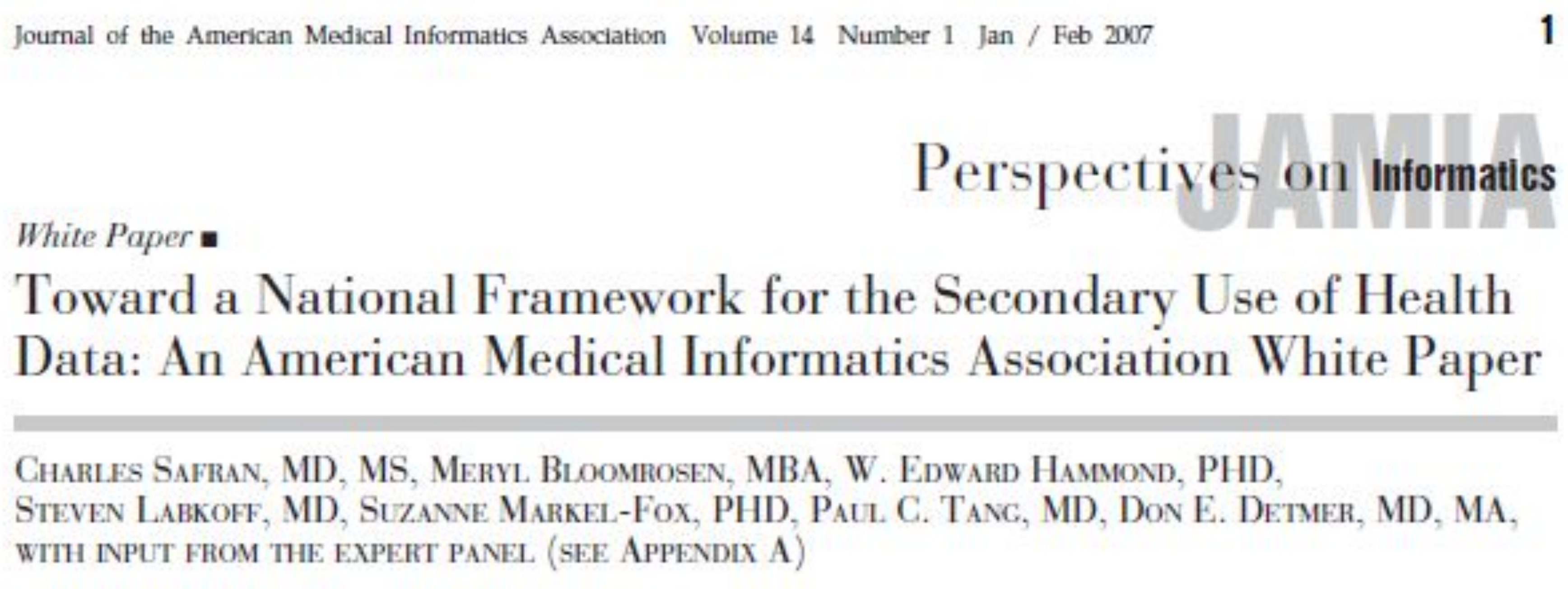
Context

- More and more data are produced daily
- **Information technologies in healthcare**
  - **Reimbursement data**
  - **Electronic Health Record (EHR)**
  - **Biology**
  - **Radiology**
- **Research data**
  - **Clinical research**
  - **Epidemiology**
- Internet of things
- Omics



Brysonwong01 [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)]

Secondary use of biomedical data



Journal of the American Medical Informatics Association Volume 14 Number 1 Jan / Feb 2007 1

JAMIA Perspectives on Informatics

White Paper ■

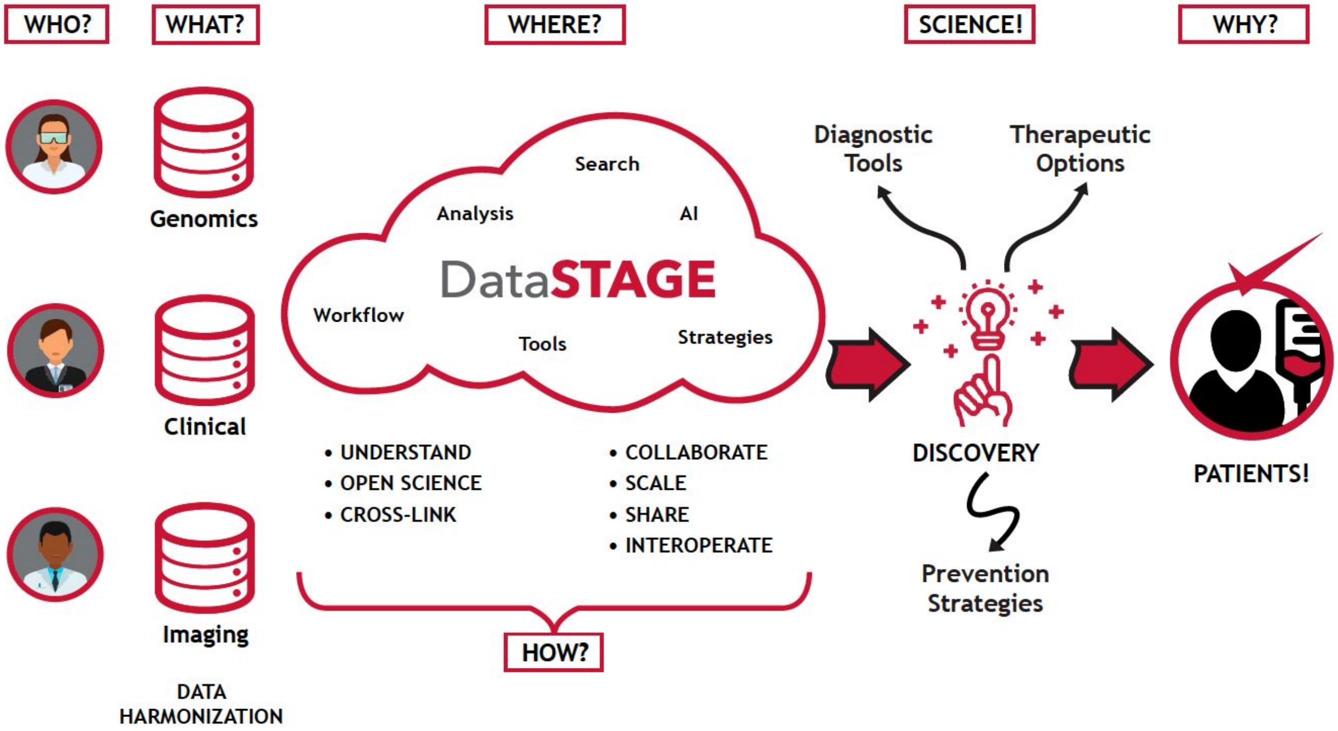
Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper

CHARLES SAFRAN, MD, MS, MERYL BLOOMROSEN, MBA, W. EDWARD HAMMOND, PHD,
STEVEN LABKOFF, MD, SUZANNE MARKEL-FOX, PHD, PAUL C. TANG, MD, DON E. DETMER, MD, MA,
WITH INPUT FROM THE EXPERT PANEL (SEE APPENDIX A)

Secondary use of biomedical data

> *“Secondary use of health data can enhance healthcare experiences for individuals, expand knowledge about disease and appropriate treatments, strengthen understanding about the effectiveness and efficiency of our healthcare systems, support public health and security goals, and aid businesses in meeting the needs of their customers”*

Reusing research data DataSTAGE

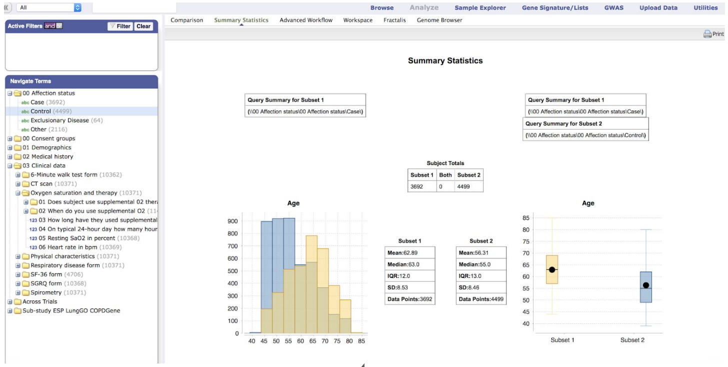


https://www.nhlbidatastage.org/

Database of Genotypes and Phenotypes (dbGaP)

- Archive and distribute NIH studies
    - Genotypes
    - Phenotypes
- Enable data retrieval
- Can not be computed directly
    - Encrypted files
    - XML dictionaries

Architecture



Summary Statistics

Query Summary for Subset 1
(\\00 Affection status\00 Affection status\Case\)

Query Summary for Subset 1
(\\00 Affection status\00 Affection status\Case\)

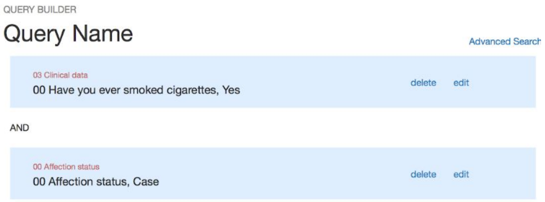
Query Summary for Subset 2
(\\00 Affection status\00 Affection status\Control\)

Subject Totals

Subset 1	Both	Subset 2
3692	0	4499

Subset 1
Mean:62.89
Median:63.0
IQR:12.0
SD:8.53
Data Points:3692

Subset 2
Mean:56.31
Median:55.0
IQR:13.0
SD:8.46
Data Points:4499



QUERY BUILDER

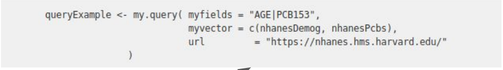
Query Name

Advanced Search

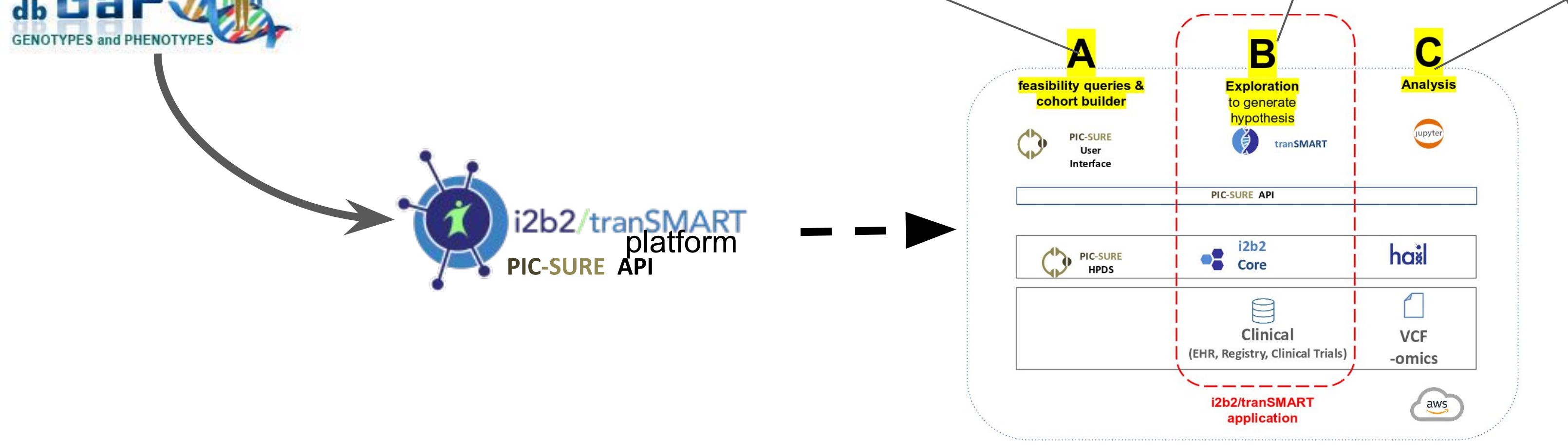
03 Clinical data
00 Have you ever smoked cigarettes, Yes

AND

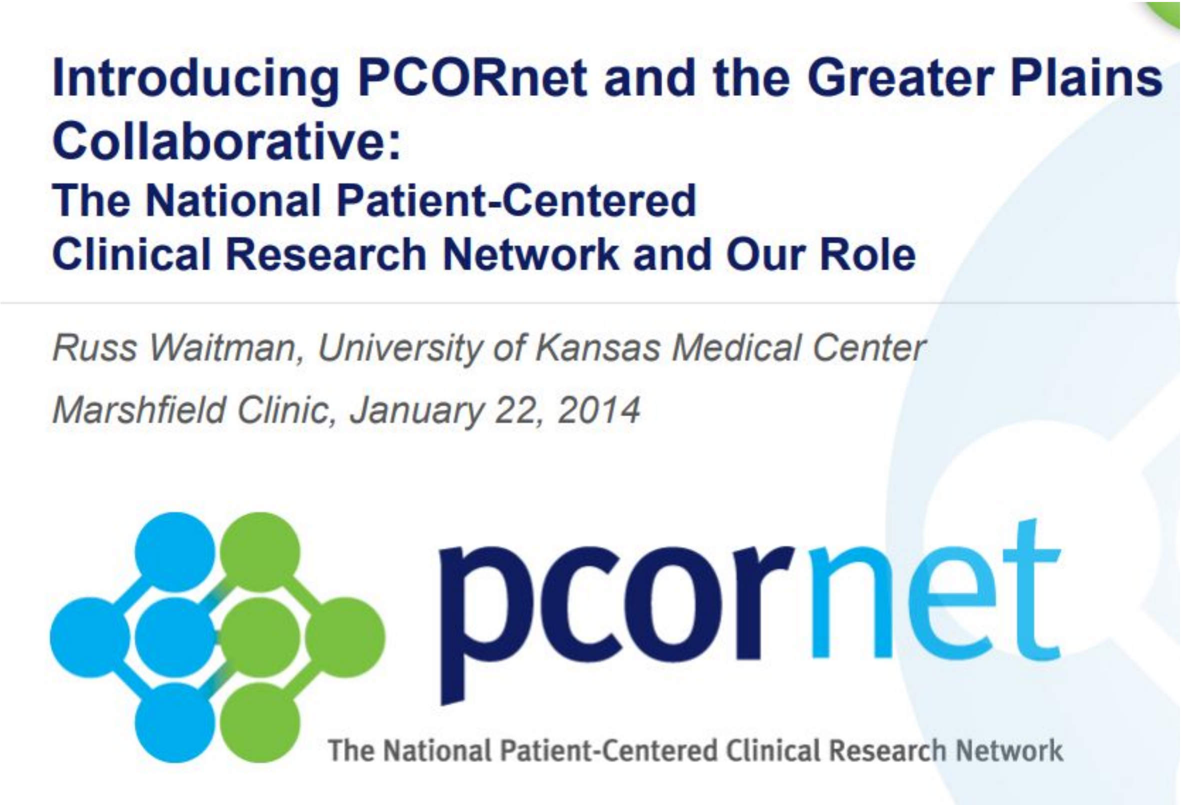
00 Affection status
00 Affection status, Case



```
queryExample <- my.query( myfields = "AGE|PCB153",
                          myvector = c(nhanesDemog, nhanesPcbs),
                          url        = "https://nhanes.hms.harvard.edu/"
          )
```



Clinical data Networks



Introducing PCORnet and the Greater Plains Collaborative:

The National Patient-Centered Clinical Research Network and Our Role

Russ Waitman, University of Kansas Medical Center

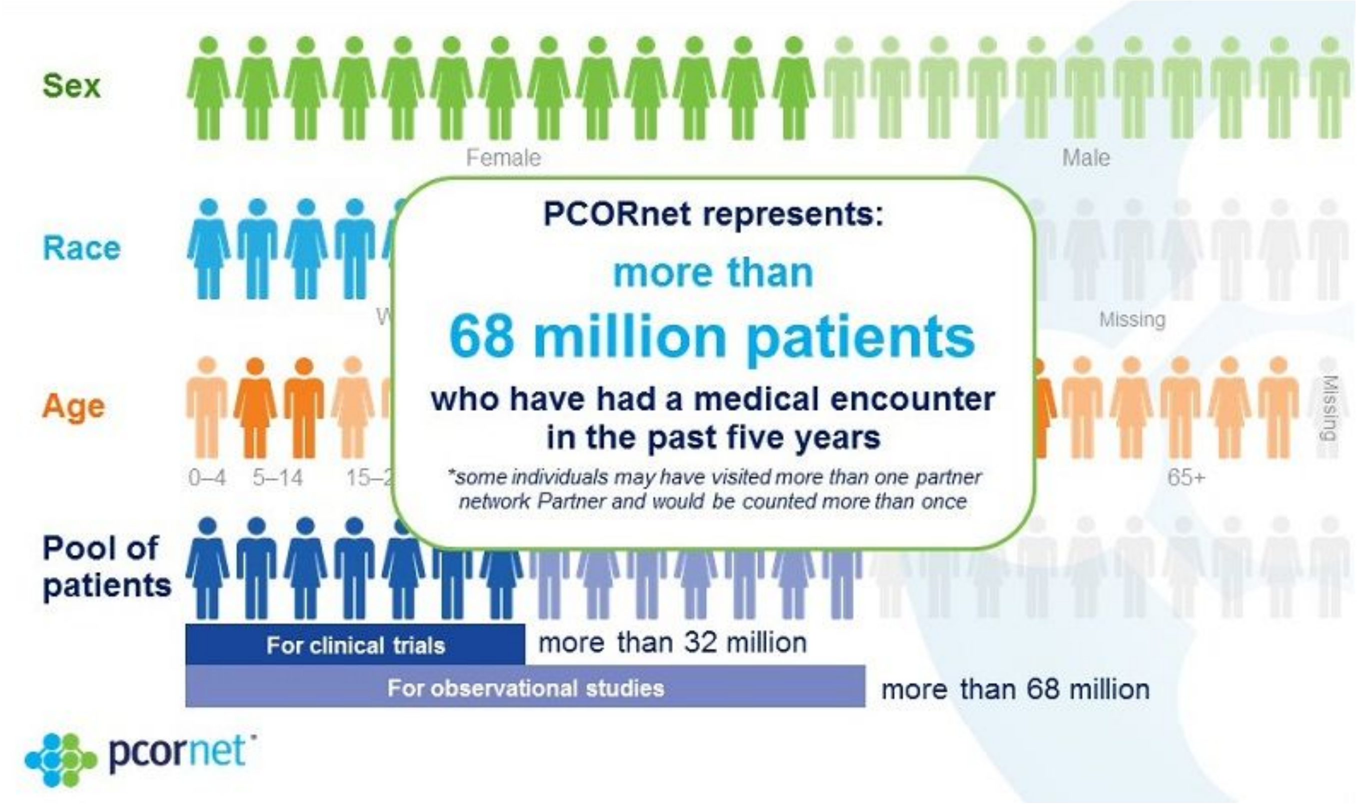
Marshfield Clinic, January 22, 2014

pcornet

The National Patient-Centered Clinical Research Network

https://pcornet.org/

Clinical data Networks



PCORnet represents:

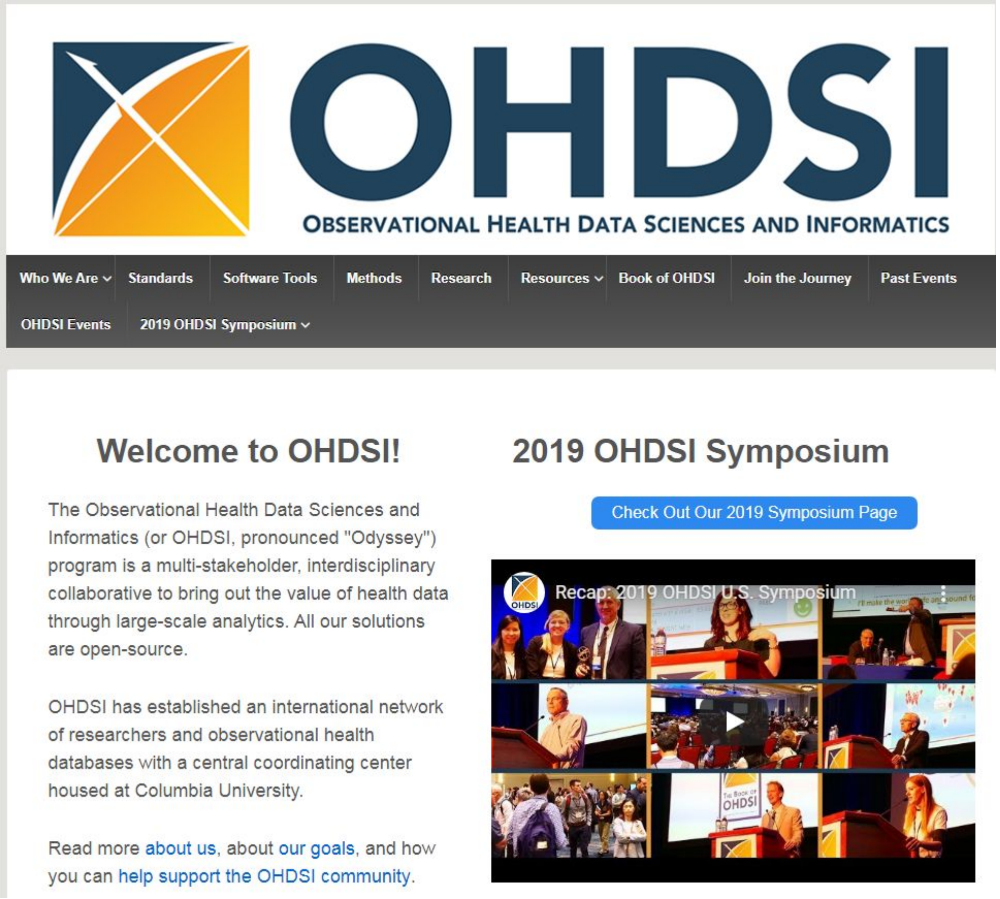
more than

68 million patients

who have had a medical encounter in the past five years

*some individuals may have visited more than one partner network Partner and would be counted more than once

Clinical data Networks



Welcome to OHDSI!

The Observational Health Data Sciences and Informatics (or OHDSI, pronounced "Odyssey") program is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics. All our solutions are open-source.

OHDSI has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University.

Read more about us, about our goals, and how you can help support the OHDSI community.

2019 OHDSI Symposium

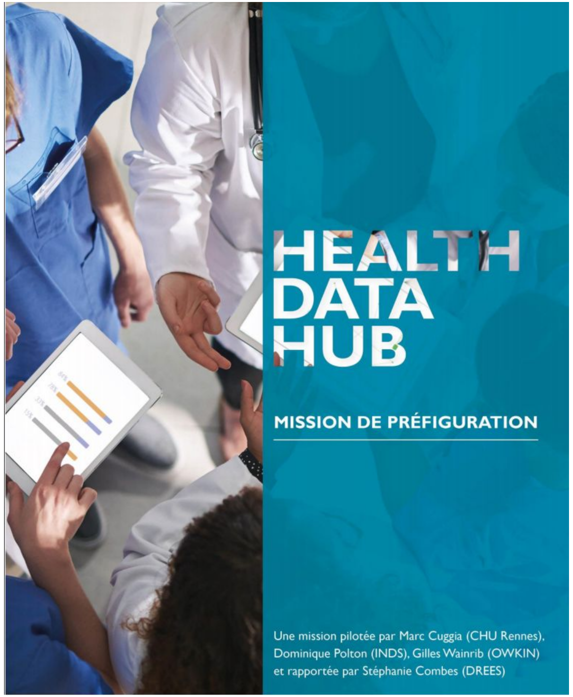
https://ohdsi.org/

Secondary use and AI...

Mission VILLANI

DONNER UN SENS
À L'INTELLIGENCE
ARTIFICIELLE

POUR UNE STRATÉGIE
NATIONALE ET EUROPÉENNE



HEALTH
DATA
HUB

MISSION DE PRÉFIGURATION

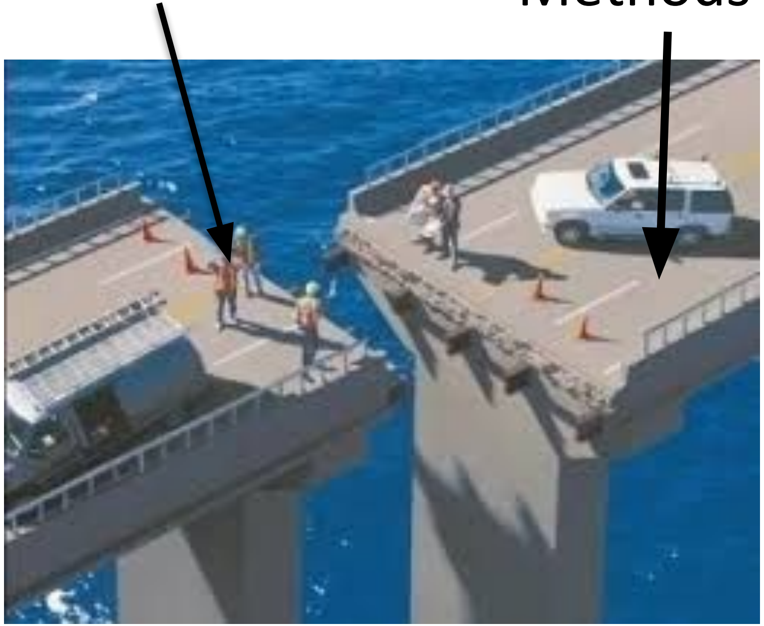
Une mission pilotée par Marc Cuggia (CHU Rennes),
Dominique Polton (INDS), Gilles Wainrib (OWKIN)
et rapportée par Stéphanie Combes (DREES)

AI is a tool... not a goal

- User need comes before method choices
    - Identify domain expert
    - Define users needs
    - Choose methods
- Evaluate➔ There is no magic !

Choose appropriate tools...

Identify user needs...



Needs

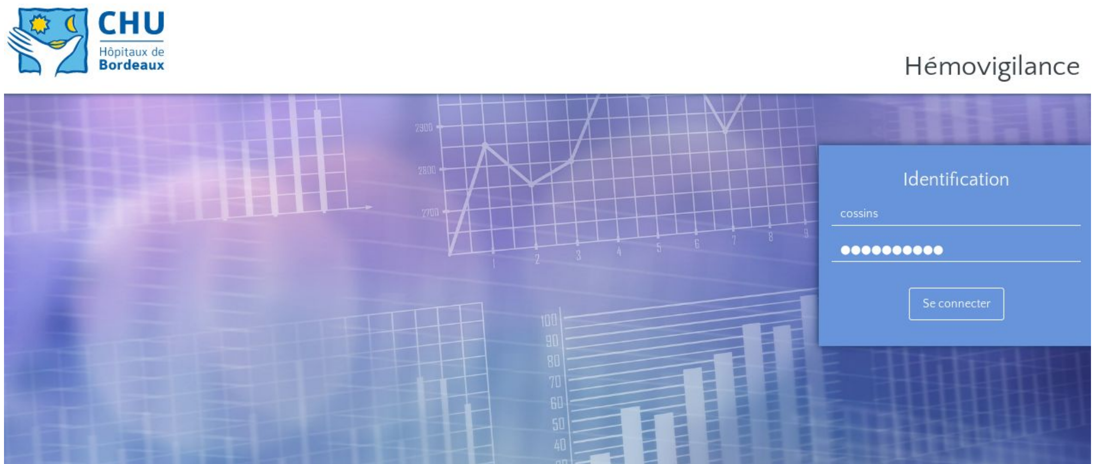
Methods

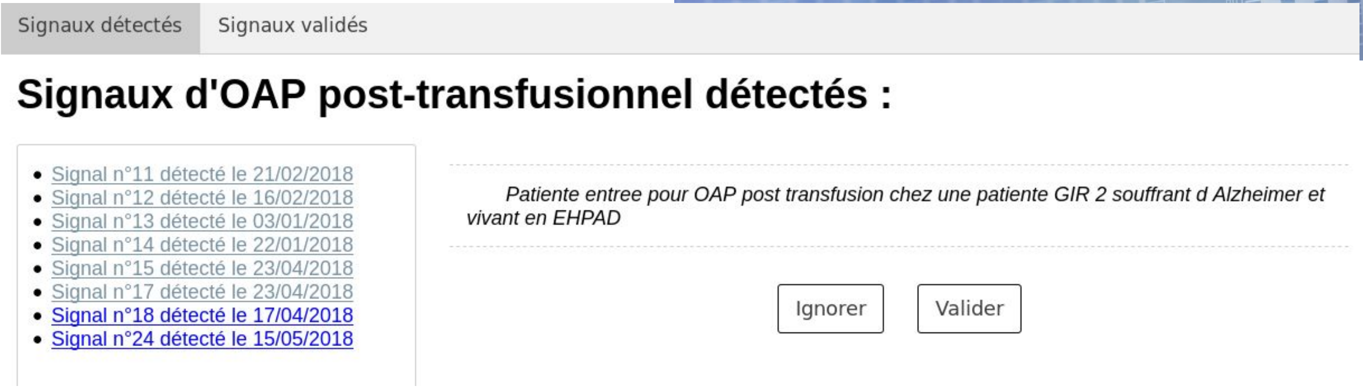
What can we do with these data ?

Acute post-transfusion pulmonary edema

- Transfusion adverse event
- Should be declared by physicians
    - Underestimated
- Methods:
    - Simple free text search for concepts co-occurrence in a single sentence
    - Manual review of the sentence (fast human filtering)
    - Contextual validation of selected cases

User Friendly UIs





Signaux d'OAP post-transfusionnel détectés :

- Signal n°11 détecté le 21/02/2018
- Signal n°12 détecté le 16/02/2018
- Signal n°13 détecté le 03/01/2018
- Signal n°14 détecté le 22/01/2018
- Signal n°15 détecté le 23/04/2018
- Signal n°17 détecté le 23/04/2018
- Signal n°18 détecté le 17/04/2018
- Signal n°24 détecté le 15/05/2018

Patiente entree pour OAP post transfusion chez une patiente GIR 2 souffrant d Alzheimer et vivant en EHPAD

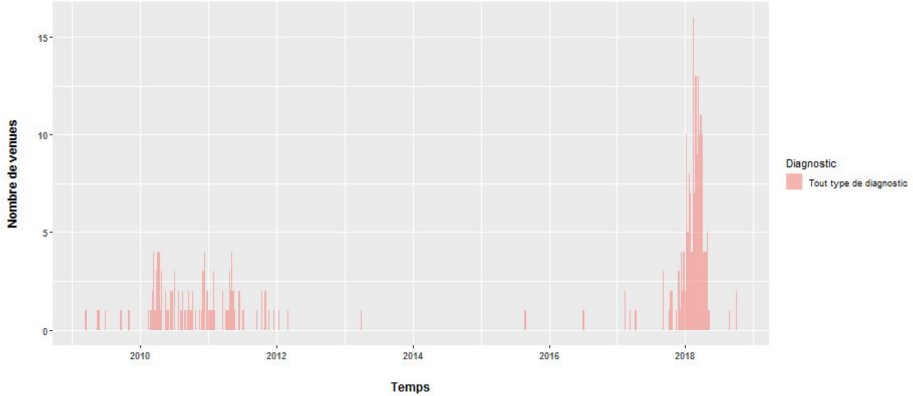
DetectTACO

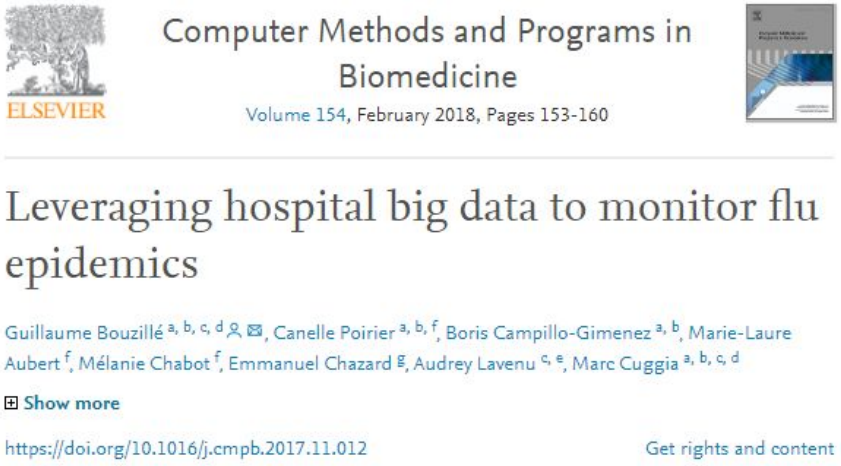
Outil de détection des OAP post-transfusionnels.

Results

- Without secondary use ⇒ 1 case identified (3 months)
- Using data
  - 102 « possible cases » detected
  - 8 cases validated
  - Cluster identification
  - Apply correction mesure

Epidemic monitoring





Computer Methods and Programs in Biomedicine

Volume 154, February 2018, Pages 153-160

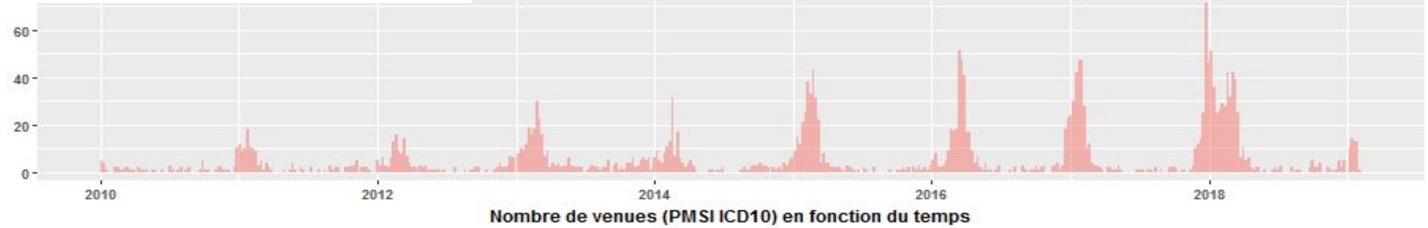
Leveraging hospital big data to monitor flu epidemics

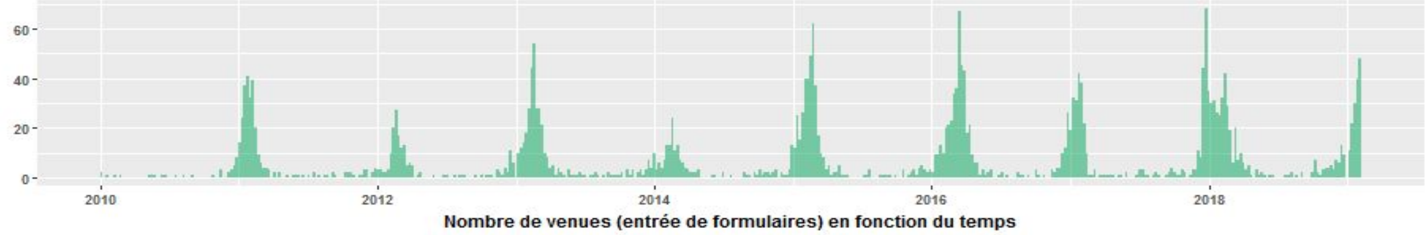
Guillaume Bouzillé [a, b, c, d], Canelle Poirier [a, b, f], Boris Campillo-Gimenez [a, b], Marie-Laure Aubert [f], Mélanie Chabot [f], Emmanuel Chazard [g], Audrey Lavenu [c, e], Marc Cuggia [a, b, c, d]

⊞ Show more

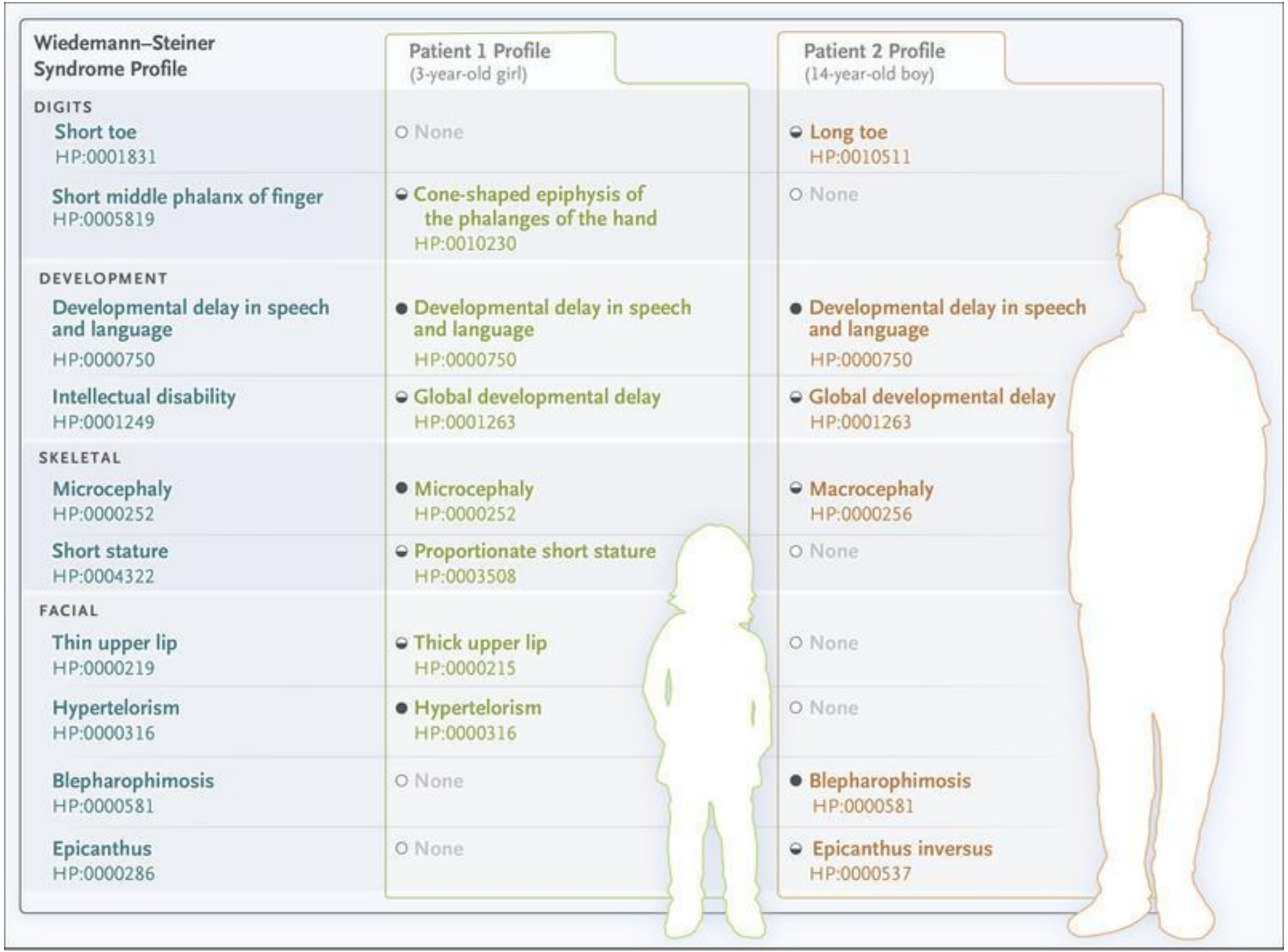
https://doi.org/10.1016/j.cmpb.2017.11.012

Get rights and content





Phenotyping



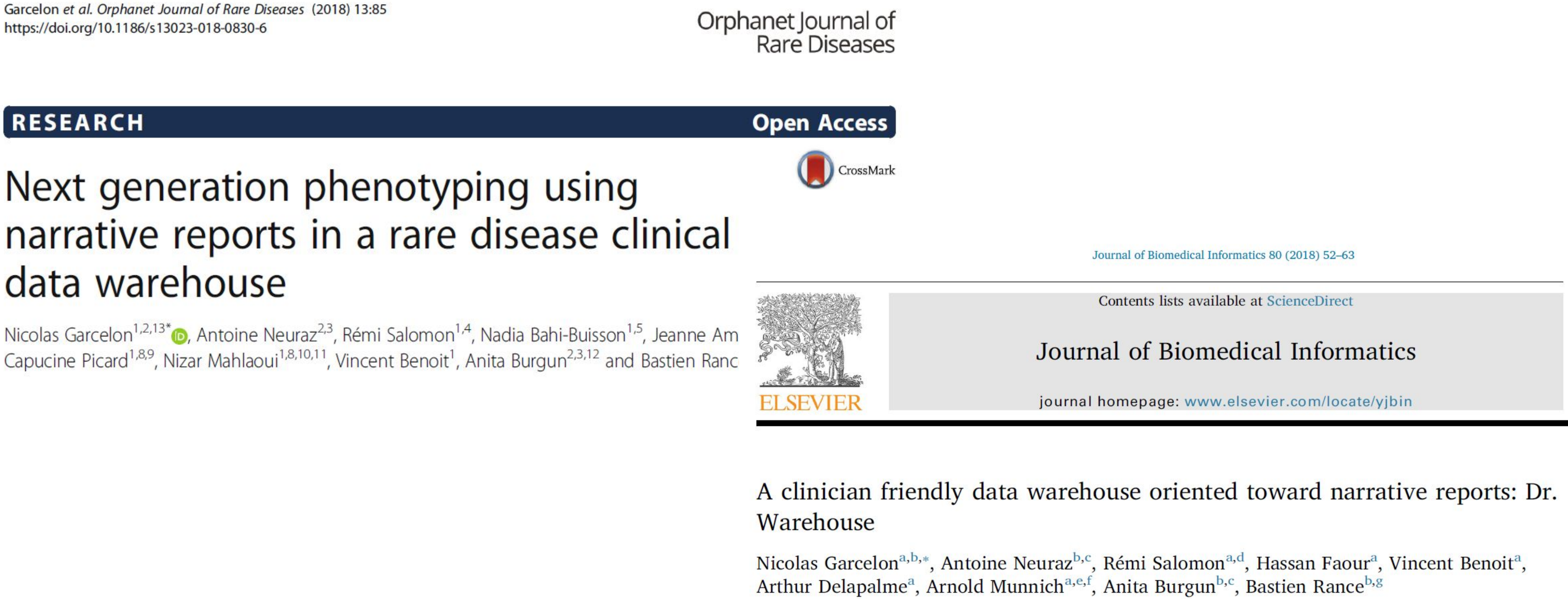
Wiedemann–Steiner Syndrome Profile	Patient 1 Profile (3-year-old girl)	Patient 2 Profile (14-year-old boy)
DIGITS		
Short toe HP:0001831	○ None	◒ Long toe HP:0010511
Short middle phalanx of finger HP:0005819	◒ Cone-shaped epiphysis of the phalanges of the hand HP:0010230	○ None
DEVELOPMENT		
Developmental delay in speech and language HP:0000750	● Developmental delay in speech and language HP:0000750	● Developmental delay in speech and language HP:0000750
Intellectual disability HP:0001249	◒ Global developmental delay HP:0001263	◒ Global developmental delay HP:0001263
SKELETAL		
Microcephaly HP:0000252	● Microcephaly HP:0000252	◒ Macrocephaly HP:0000256
Short stature HP:0004322	◒ Proportionate short stature HP:0003508	○ None
FACIAL		
Thin upper lip HP:0000219	◒ Thick upper lip HP:0000215	○ None
Hypertelorism HP:0000316	● Hypertelorism HP:0000316	○ None
Blepharophimosis HP:0000581	○ None	● Blepharophimosis HP:0000581
Epicanthus HP:0000286	○ None	◒ Epicanthus inversus HP:0000537

Haendel, M.A., Chute, C.G., Robinson, P.N., 2018. Classification, Ontology, and Precision Medicine. N. Engl. J. Med. 379, 1452–1462. https://doi.org/10.1056/NEJMra1615014

Phenotyping

Dr Warehouse

Plus proche de vos patients



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

CrossMark

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 Am
Capucine Picard[1,8,9], Nizar Mahlaoui[1,8,10,11], Vincent Benoit[1], Anita Burgun[2,3,12] and Bastien Ranc

Journal of Biomedical Informatics 80 (2018) 52–63

ELSEVIER

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 Munnich[a,e,f], Anita Burgun[b,c], Bastien Rance[b,g]

The NEW ENGLAND JOURNAL of MEDICINE

REVIEW ARTICLE

Elizabeth G. Phimister, Ph.D., *Editor*

Classification, Ontology, and Precision Medicine

Melissa A. Haendel, Ph.D., Christopher G. Chute, M.D., Dr.P.H.,
and Peter N. Robinson, M.D.

N ENGL J MED 379;15 NEJM.ORG OCTOBER 11, 2018

Need for semantic standardisation

- Identify information meaning uniquely
- Extract concepts from free text
- Organise information depending on its meaning
  - Medication
  - Phenotypes
  - Disease
- Drive data Visualization through external knowledge
- Based on Semantic resources
  - Terminologies
  - Ontologies

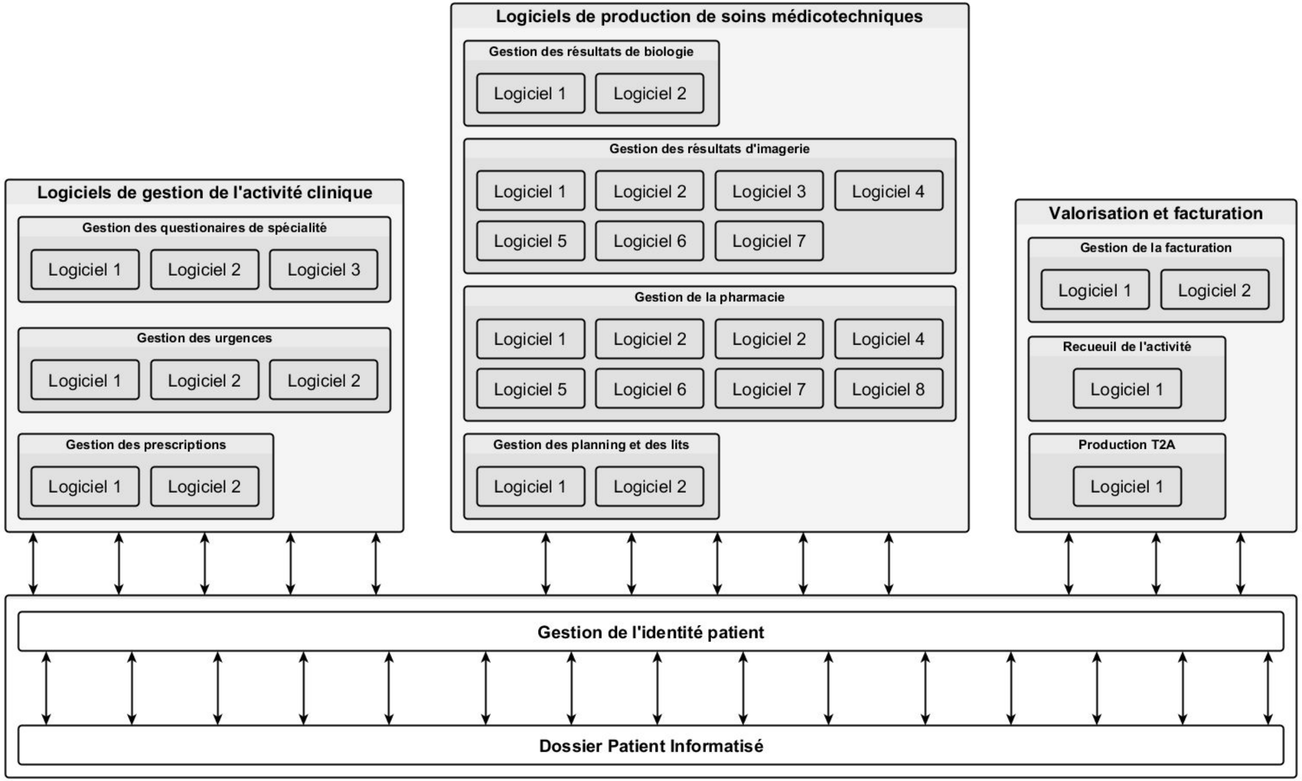
Re-using hospital data

Hospital Information systems

- Multiple distributed medical applications
  - Electronic Health record
  - Administrative data
  - Radiology
  - Biology
  - Etc…
- Build for healthcare purpose (not for secondary use)
- Reimbursement data production

Lead to data silos

Hospital Information Systems



Logiciels de gestion de l'activité clinique

Gestion des questionaires de spécialité
- Logiciel 1
- Logiciel 2
- Logiciel 3

Gestion des urgences
- Logiciel 1
- Logiciel 2
- Logiciel 2

Gestion des prescriptions
- Logiciel 1
- Logiciel 2

Logiciels de production de soins médicotechniques

Gestion des résultats de biologie
- Logiciel 1
- Logiciel 2

Gestion des résultats d'imagerie
- Logiciel 1
- Logiciel 2
- Logiciel 3
- Logiciel 4
- Logiciel 5
- Logiciel 6
- Logiciel 7

Gestion de la pharmacie
- Logiciel 1
- Logiciel 2
- Logiciel 2
- Logiciel 4
- Logiciel 5
- Logiciel 6
- Logiciel 7
- Logiciel 8

Gestion des planning et des lits
- Logiciel 1
- Logiciel 2

Valorisation et facturation

Gestion de la facturation
- Logiciel 1
- Logiciel 2

Recueuil de l'activité
- Logiciel 1

Production T2A
- Logiciel 1

Gestion de l'identité patient

Dossier Patient Informatisé

Interoperability

- Standards (IHE, HL7, CDISC)
  - Messages (HPRIM, HL7)
  - Format
- Information exchange
- Depends on vendors implementation +++
- Complex to implement

Recently standard HL7 / FHIR

https://www.hl7.org/fhir/

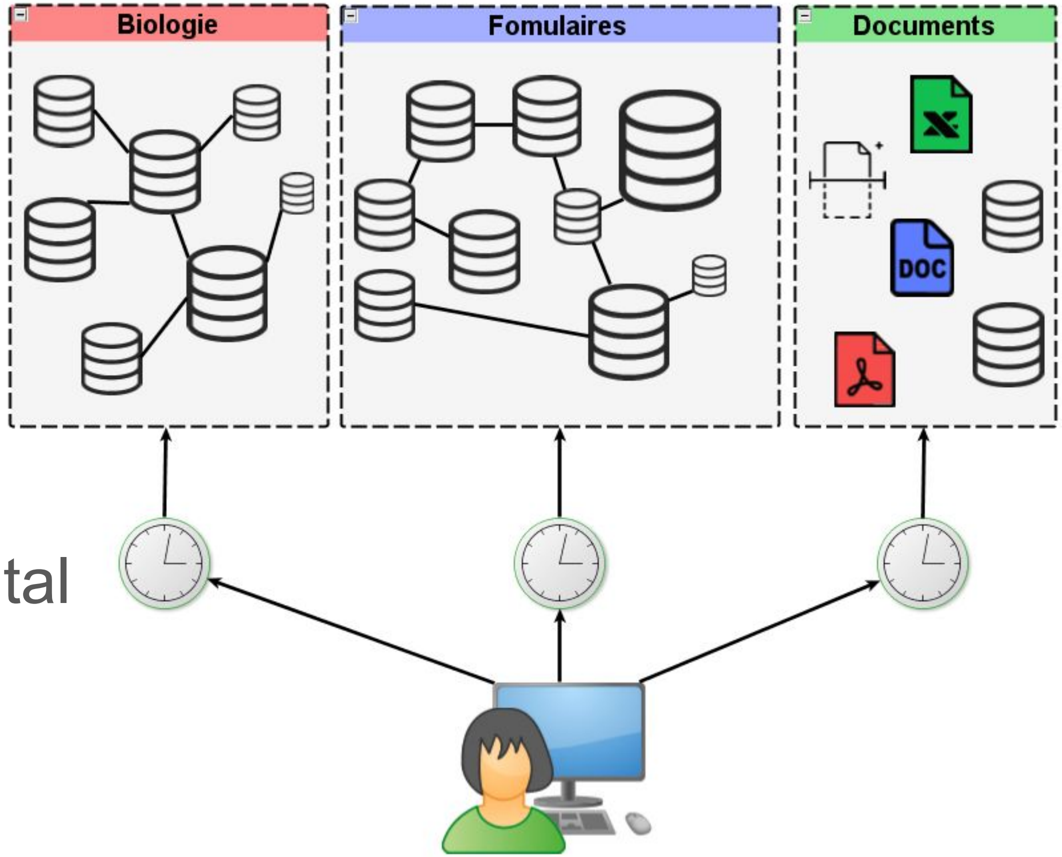
Issues

Separated data (silos)

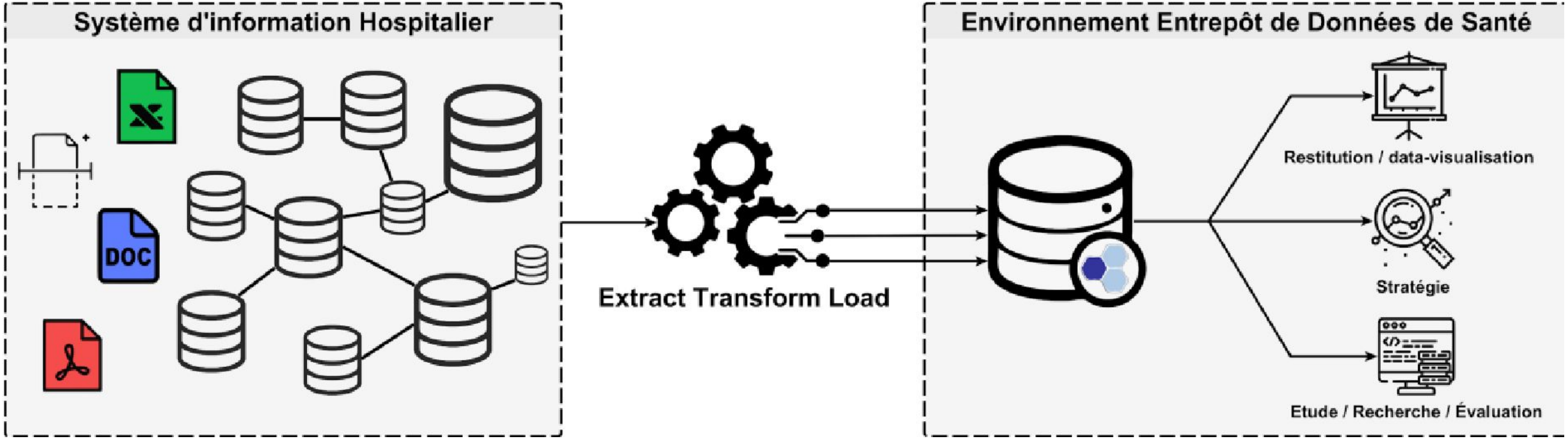
Data Heterogeneity

Bordeaux University Hospital

- 100+ Applications
- 10 000+ Tables



Clinical data warehouse (CDW)


Système d'information Hospitalier
DOC
Extract Transform Load
Environnement Entrepôt de Données de Santé
Restitution / data-visualisation
Stratégie
Etude / Recherche / Évaluation

ETL - Extract

- Extract data from Hospital Information System
  - Need knowledge of data model
  - This usually needs to be reverse engineered
- Another solution
  - Use exchange standard
  - Leverage standardised messages
  - Necessitate to listen data flows in real time
  - FHIR may offer easier solution for data extraction

ETL - Transform

Transform data from HIS

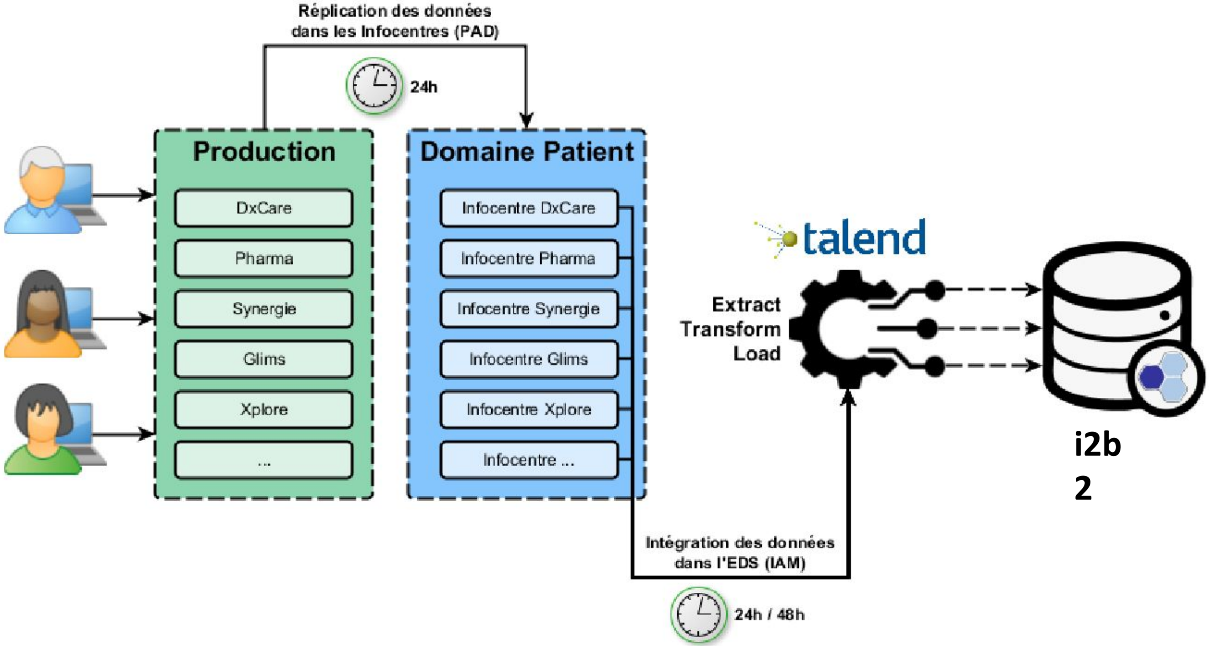
- Technical transformation
  - Relational model to CDW information model
- Semantic transformation
  - Harmonize terminologies
  - Annotate information (meta modeling)
  - Leverage Ontologies and terminologies
  - May benefit from multi terminology server

ETL - Load

Load data into the EDS information model

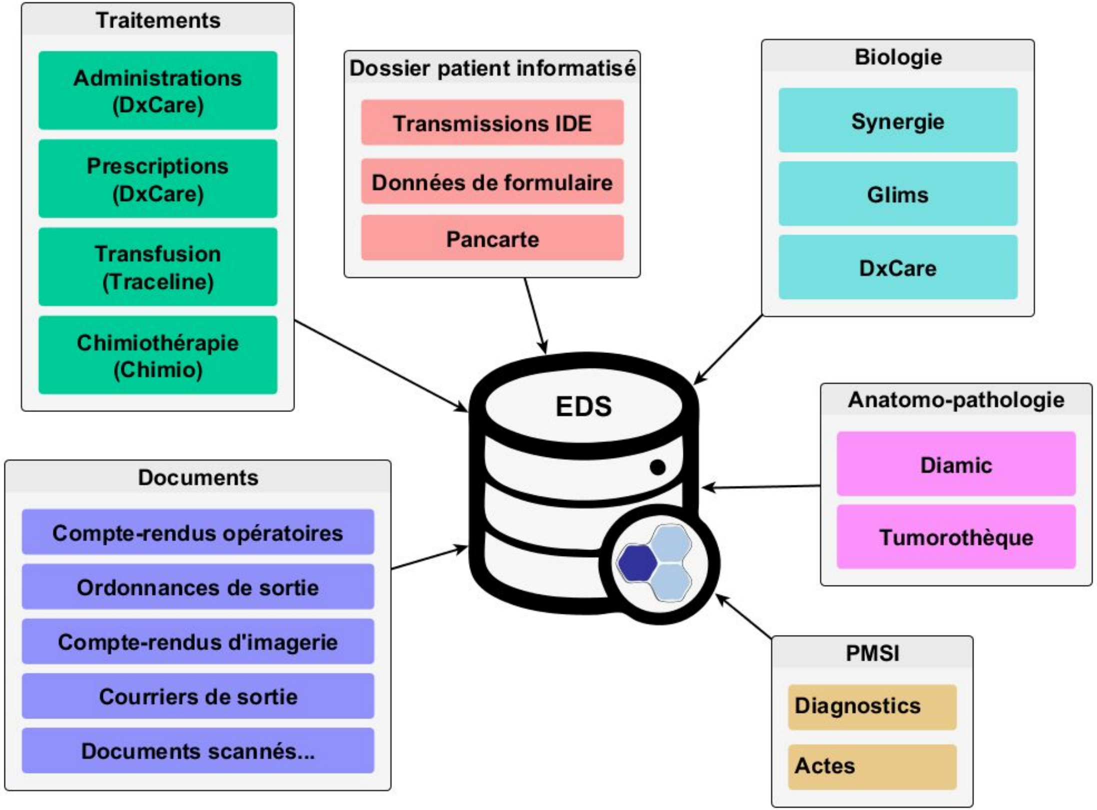
- Trade of
    - Update data (add, remove, update)
    - Drop and replace

Example Bordeaux University Hospital CDW

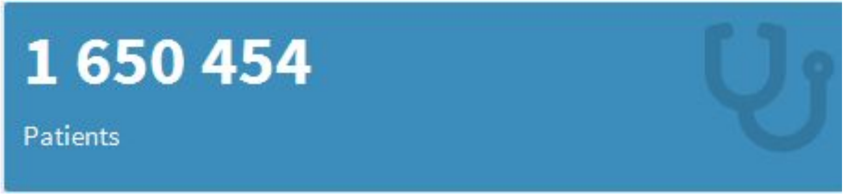


i2b
2

CDW: data integrated

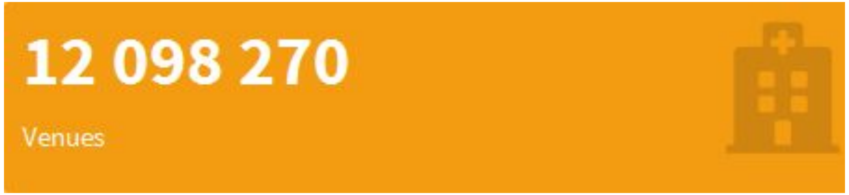


CDW: data integrated



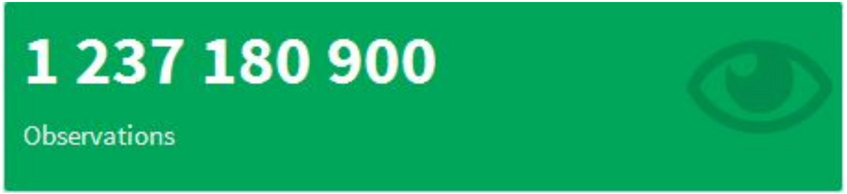
1 650 454

Patients



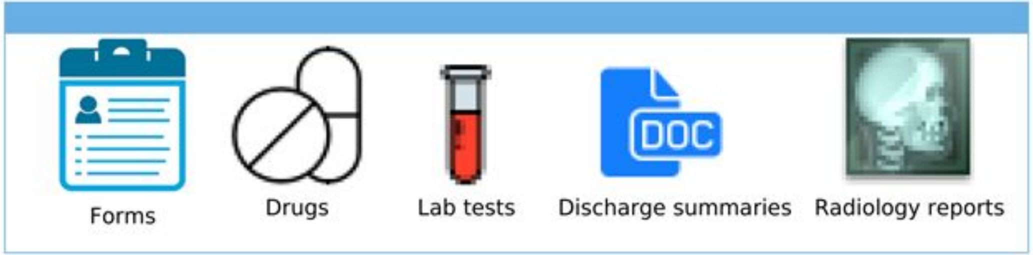
12 098 270

Venues



1 237 180 900

Observations



Data access and data privacy

Driving Principles

- Security et traceability
  - Data are processed in situ (move algorithm not data).
- Transparency
  - Patient information (General Data Protection Regulation)
  - Open source / Open data / open science

- Optimise data availability for secondary use and data privacy

Confidentiality, integrity, availability

- Trade of between confidentiality, integrity, availability
- A system with high security will limit availability
  - Highest security ⇒ offline storage (not accessible at all !!!)
- A system with high availability may cause privacy breach and data leak
  - The most available ⇒ Open data available for all (no authorization, no authentication!!!)
- High integrity is time consuming
  - Perfect and complete data is unreachable (data will never be available)

Data privacy risk levels

4 levels

1. Dictionaries (biologie, formulaire etc…)
2. Category counts
3. Query with patient numbers (obfuscation)
4. Detailed data
    a. Identifiers (HIPAA - Health Insurance Portability and Accountability Act),
    b. Pseudonymised

Goals of data processing

- Within structure scope
  - Health care quality, vigilances, evaluation ...
- Research, studies
  - Clinical research
  - Epidemiology
  - Research databases
  - Biobanks

What can be done...

	Routine activity	Research, studies
Level 1 (dict)		
Level 2 (cat counts)		
Level 3 (counts)		
Level 4 (detailed)		

What can be done...

	Routine activity	Research, studies
Level 1 (dict)	Open data	Open data
Level 2 (cat counts)		
Level 3 (counts)		
Level 4 (detailed)		

What can be done...

	Routine activity	Research, studies
Level 1 (dict)	Open data	Open data
Level 2 (cat counts)	Open data	Open data
Level 3 (counts)		
Level 4 (detailed)		

Co-occurrence matrix (Natural language processing and structured data)

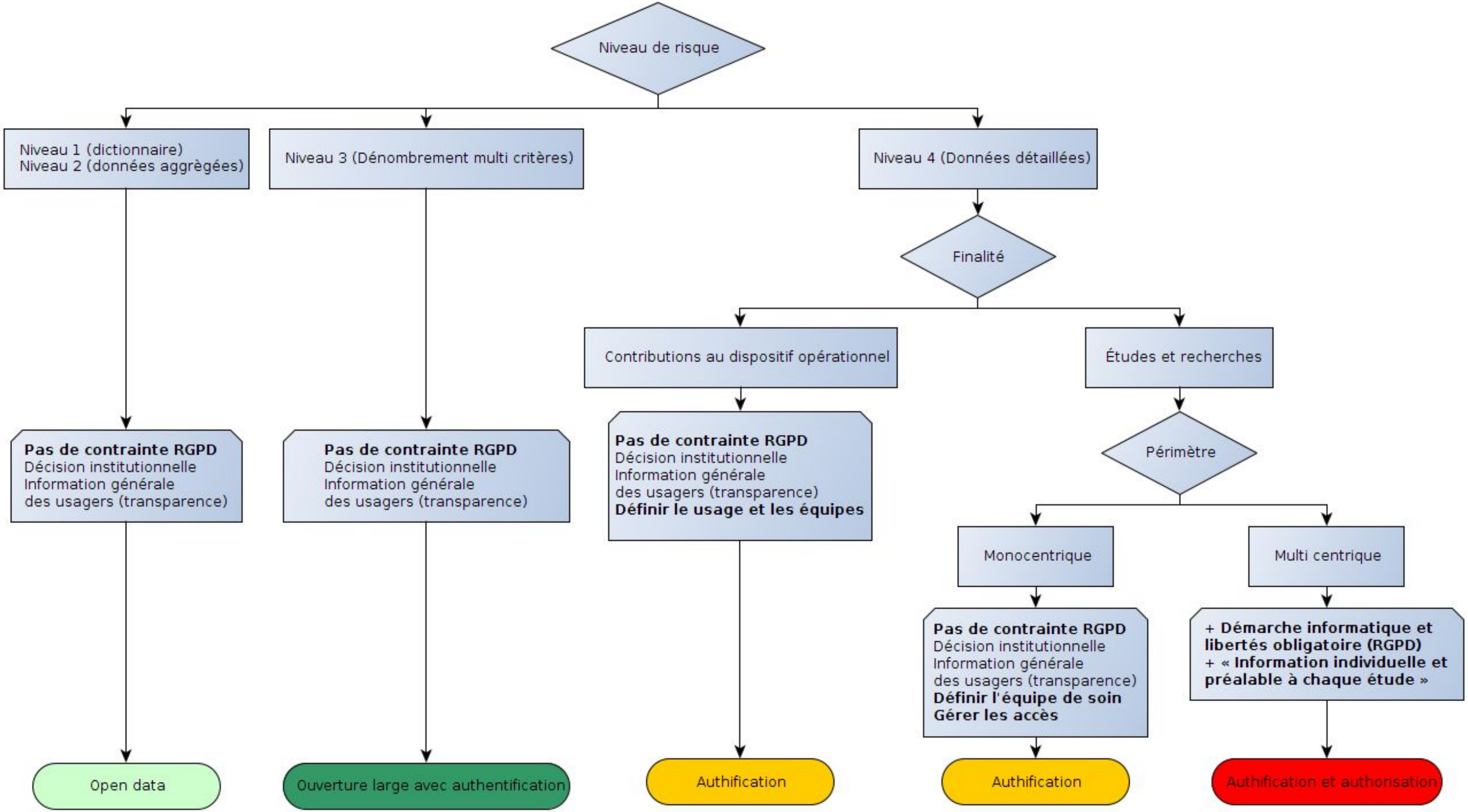
- Enables embeddings
- Train models

What can be done...

	Routine activity	Research, studies
Level 1 (dict)	Open data	Open data
Level 2 (cat counts)	Open data	Open data
Level 3 (counts)	Authenticated user / widely available	Authenticated user / widely available
Level 4 (detailed)		

What can be done...

	Routine activity	Research, studies
Level 1 (dict)	Open data	Open data
Level 2 (cat counts)	Open data	Open data
Level 3 (counts)	Authenticated user / widely available	Authenticated user / widely available
Level 4 (detailed)	Authorized user / specific missions	Authorized user / Patient consent


Niveau de risque
Niveau 1 (dictionnaire)
Niveau 2 (données aggrègées)
Niveau 3 (Dénombrement multi critères)
Niveau 4 (Données détaillées)
Finalité
Contributions au dispositif opérationnel
Études et recherches
Périmètre
Monocentrique
Multi centrique
Pas de contrainte RGPD
Décision institutionnelle
Information générale
des usagers (transparence)
Pas de contrainte RGPD
Décision institutionnelle
Information générale
des usagers (transparence)
Pas de contrainte RGPD
Décision institutionnelle
Information générale
des usagers (transparence)
Définir le usage et les équipes
Pas de contrainte RGPD
Décision institutionnelle
Information générale
des usagers (transparence)
Définir l'équipe de soin
Gérer les accès
+ Démarche informatique et libertés obligatoire (RGPD)
+ « Information individuelle et préalable à chaque étude »
Open data
Ouverture large avec authentification
Authification
Authification
Authification et authorisation