Utilisation des entrepôts de données pour la recherche

Expérience de l'HEGP

Gilles Chatellier

INSERM CIC 1418 , URC HEGP- AP-HP

Et Faculté Paris Descartes

Types de bases de données

- **Large cohortes**
  - Programme REIN (**R**éseau épidémiologique et **I**nformation en **N**éphrologie
- **Bases de données médico-administratives**
  - SNIIR-AM
  - EGB
- **Entrepôts de données**
  - AP-HP
  - HEGP
  - Autres CHU

- **Essais cliniques randomisés (en développement)**

Utilisation des bases de données

Aide au montage d'études cliniques

- Prévalence/incidence (puissance)
- Identification de centres

Etudes originales

- Pharmacoépidémiologie
- Accès aux soins
- Qualité des soins et des pratiques
- Etudes cas-témoins

Extraction des données/organisation du fichier de travail

- Comprendre la structure de la base de données
- Préparation d'un plan d'analyse détaillé: données multiples, horizon temporel, traitement des données manquantes …
- Extraction des données requises (parfois complexe: double extraction par 2 ingénieurs indépendants…)
- Application de techniques de couplage appropriées des bases en cas de multiples sources de données
- Organisation (structure) finale de la base de travail et stockage approprié

Grandes bases de données
Limitations & Challenges

- **Par nature, non expérimentales**
- **On ne maitrise pas la qualité des données**
- **La cause du recueil des données doit être connue**
- **Imposent le recours à des connaissances spécifiques**
  - **Sur les données (PMSI)**
  - **En informatique**
  - **En biostatistique**
- **Ne suppriment pas la bureaucratie!!**
  - **Projets écrits**
  - **IRB / CPP / CNIL dans le nouveau contexte initié par la loi Jardé**

Development and Validation of Algorithms to Identify Statin Intolerance in a US Administrative Database

Table 2 – Definition of statin-related AEs.

AE	Clinical term	ICD-9-CM code
Musculoskeletal AEs/side effects		
Myalgia and/or myositis	Muscle pain, spasm, weakness, discomfort, soreness, cramps, or aching	728.85 (spasm of muscle); 729.82 (cramp in limb); 728.87 (muscle weakness-generalized); 729.1x (myalgia and myositis, unspecified)
Arthralgia	Joint pain, joint stiffness	719.4x (pain in joint); 719.5x (stiffness in joint)
Rhabdomyolysis	Rhabdomyolysis	728.88 (rhabdomyolysis)
Myopathy	Myopathy, toxic myopathy	359.4 (toxic myopathy); 359.89 (other myopathy); 359.9 (myopathy, unspecified)
Other muscle toxicity	Limb pain, ligament pain, fascia pain, limb discomfort, other drug poisoning	728.9x (unspecified disorder of muscle, ligament, and fascia); 729.1x (pain in limb); E980.4 (injury from other specified drugs)
Elevated CPK	CPK >5× ULN	
Gastrointestinal-related AEs/side effects		
Nausea	Nausea, vomiting	787.0 (nausea and/or vomiting)
Constipation	Constipation	564.0 (constipation)
Diarrhea	Diarrhea	564.5 (functional diarrhea); 787.91 (diarrhea)
Other gastrointestinal distress	Flatulence, gas, bloating, abdominal pain, gastritis, duodenitis	787.3 (flatulence, eructation, and gas pain); 789.0 (abdominal pain); 535 (gastritis and duodenitis)
Other AEs/side effects		
Anaphylaxis	Anaphylactic shock	995.0 (other anaphylactic shock)
Rash/facial flushing	Urticaria, angioedema, rash, hives, dermatitis, edema, erythema, rosacea, facial flushing, pruritus	782.1 (rash and other nonspecific skin eruption); 708.0 (allergic urticaria); 693.0 (dermatitis due to drugs/medicines); 995.1 (angioneurotic edema); 995.2 (other and unspecified AE of drug, medicinal, and biological substance (due) to correct medicinal substance properly administered); 695.1 (erythema multiforme); 698 (pruritus and related conditions); 695.3 (rosacea)
Cognitive impairment	Memory loss, forgetfulness, confusion	331.83 (mild cognitive impairment, so stated); 780.93 (memory loss); 799.51 (attention or concentration deficit)
Elevated LFT	LFT >3× ULN	

Note. Intolerance definition: The primary analysis included the complete list of potential Aes in the definition of SI, whereas the secondary analysis included only musculo-skeletal events.

AEs, adverse-effects; CPK, creatine-phosphokinase; ICD-9CM, International Classification of Diseases, 9th Revision, Clinical Modification; LFT, liver function test; ULN, upper limit of normal.

Kathy L. Schulman. Value In Health, 2016.

Development and Validation of Algorithms to Identify Statin Intolerance in a US Administrative Database

Table 2 – Definition of statin-related AEs.

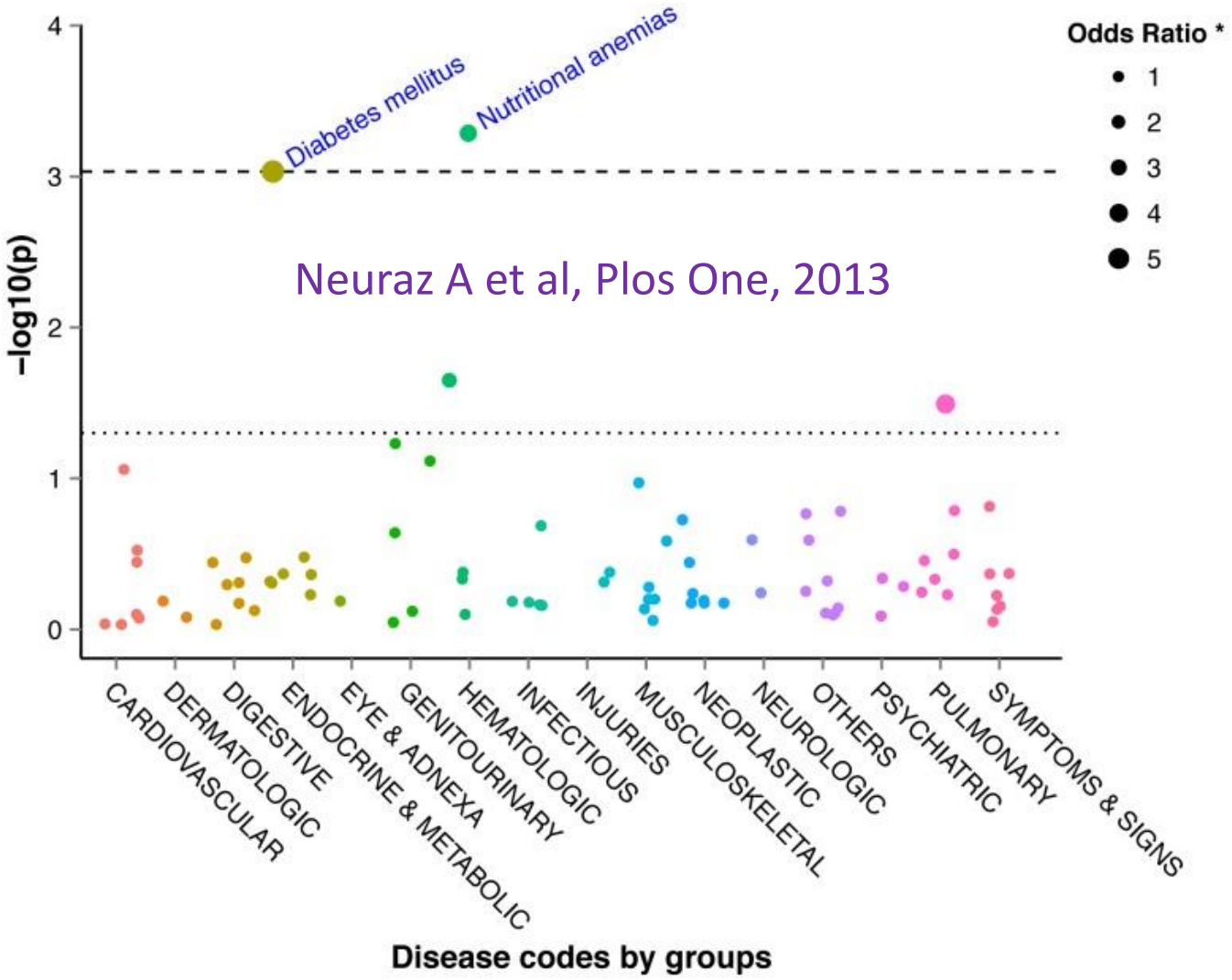
AE	Clinical term	*ICD-9-CM* code
Musculoskeletal AEs/side effects		
Myalgia and/or myositis	Muscle pain, spasm, weakness, discomfort, soreness, cramps, or aching	728.85 (spasm of muscle); 729.82 (cramp in limb); 728.87 (muscle weakness-generalized); 729.1x (myalgia and myositis, unspecified)
Arthralgia	Joint pain, joint stiffness	719.4x (pain in joint); 719.5x (stiffness in joint)
Rhabdomyolysis	Rhabdomyolysis	728.88 (rhabdomyolysis)
Myopathy	Myopathy, toxic myopathy	359.4 (toxic myopathy); 359.89 (other myopathy); 359.9 (myopathy, unspecified)
Other muscle toxicity	Limb pain, ligament pain, fascia pain, limb discomfort, other drug poisoning	728.9x (unspecified disorder of muscle, ligament, and fascia); 729.1x (pain in limb); E980.4 (injury from other specified drugs)

Kathy L. Schulman. Value In Health, 2016.

Alzheimer et EGB
Comment identifier les patients ?

- Trois modes d'identification des patients>60 ans ayant un Alzheimer ont été utilisés : l'affection de longue durée 15 (ALD15), la consommation de médicaments spécifiques de l' Alzheimer (MED) et les codages PMSI des syndromes démentiels (PMSI).
- 3802 bénéficiaires répartis en 7 populations disjointes ont été identifiés :
  - ALD15 (17.3%), MED (21.4%), PMSI (14.4%)
  - et les différentes combinaisons de ces 3 critères: ALD15+MED (27.1%), ALD15+PMSI (2.5%), PMSI+MED (6.8%) et ALD15+MED+PMSI (9.7%)

E Lenain et al. EGB; Echantillon au 1/97eme du SNIIR-AM



PHE-WAS association study

Manhattan plot of −log10 (P-values) for the 256 ICD-10 based aggregated codes between very patients with high TPMT (thiopurine *S*-methyltransferase) activity and other patients. Dotted line: P-value of 0.05 ; Dashed line: FDR corrected level of significance (q = 0.2).

Épidémiologie clinique : apport des données d'un entrepôt de données cliniques (EDC)

- Objectifs :
  - Confirmer association dysnatrémies borderlines **([130;135)&(145;150]mmol/L)** et mortalité hospitalière grâce aux données de l'EDC-HEGP
  - Identifier de nouveau facteurs de confusion grâce aux études PheWas et à l'EDC
- PheWas :
  - Association entre variables d'intérêt (mortalité, dysnatrémie) et phénotypes (ensemble des codes CIM-10), sans a priori
  - Résultats :
    - 13 codes CIM10 associés hyponatrémie borderline & mortalité
    - 6 codes CIM10 associés hypernatrémie borderline & mortalité

(Entrepôt HEGP soumis)

Épidémiologie clinique : apport des données d'un entrepôt de données cliniques (EDC)

Association Between Borderline Hyponatremia and Mortality			
	OR (IC95%)	**P**	**AIC**
Classical Model [a]	1.98 (1.73;2.68)	<.001	8098.4
Phewas Model [b]	2.59 (2.28;2.94)	<.001	11002
Final Model [c]	1.57 (1.35;1.81)	<.001	7585.5
Association Between Borderline Hypernatremia and Mortality			
	OR (IC95%)	**P**	**AIC**
Classical Model [d]	3.72(1.53;8.45)	<.001	6077.9
Phewas Model [e]	6.23(4.60;8.33)	<.001	8761.9
Final Model [f]	3.47(2.43;4.90)	<.001	5920.8

Confounding Factors Retained in the different models:

[a] Classical model: age, duration of hospital stay, number of ICD-10 codes, hospital admissions via the emergency department, ICU stay, dialysis, palliative care, Charlson Comorbidity Index

[b] PheWas model: A41, I20, I25, I48, I71, J15, J80, J96, K65, R07, R57, Z48, Z51

[c] Final model: classical model + I20, I25, I48, J80, R57, Z48, Z51

[d] Classical model: age, duration of hospital stay, number of ICD-10 codes, hospital admissions via the emergency department, ICU stay, dialysis, palliative care, Charlson Comorbidity Index

[e] PheWas model: J69, J80, J96, N17, R57, S06

[f] Final model: classical model + J69, J80, R57, S06

(Entrepôt HEGP soumis)

Timing in initiating lung cancer treatment after bronchoscopy in France: Study from medico-administrative database.

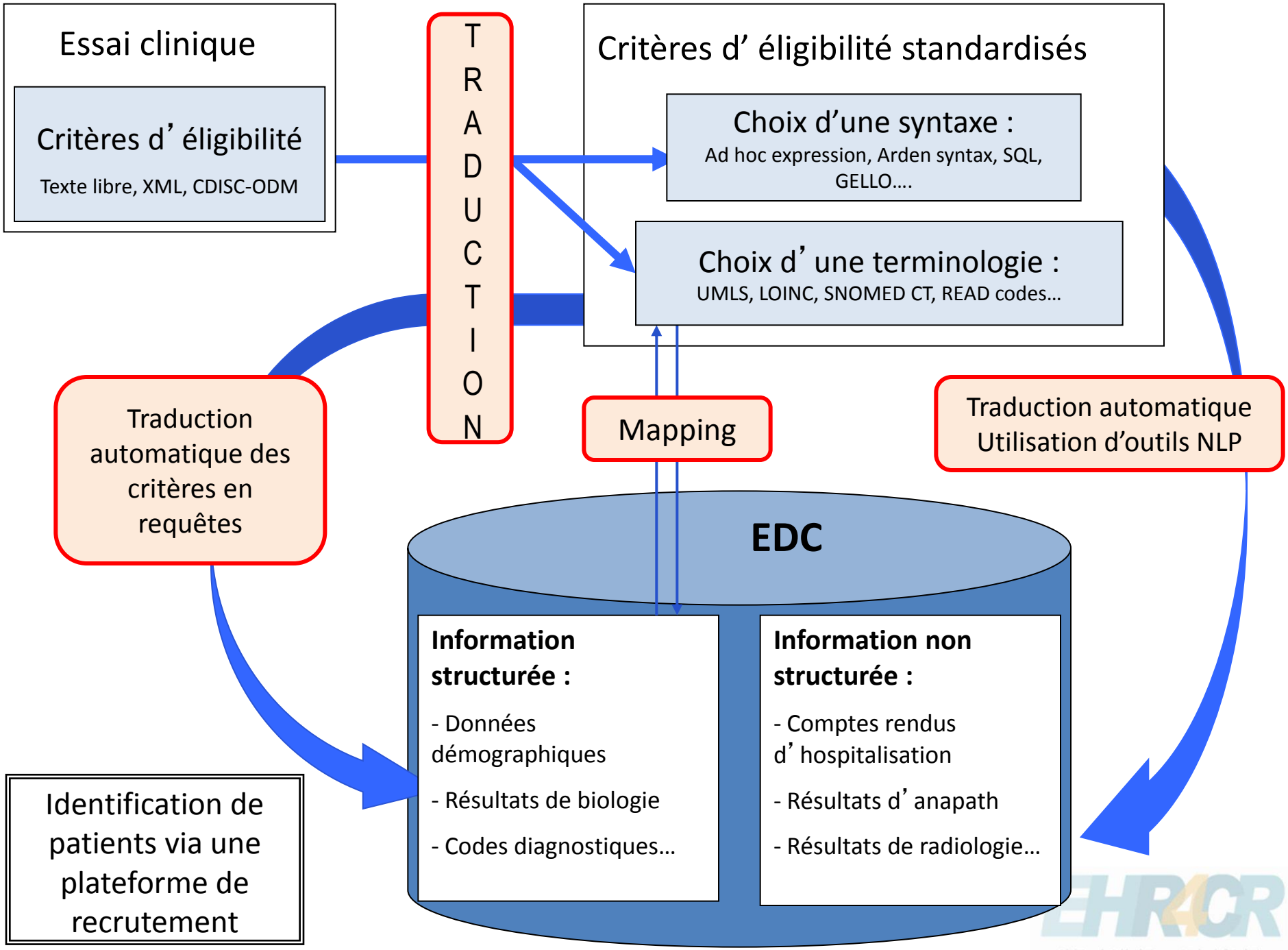
Patients issued from national hospital discharge database with a newly diagnosed Lung cancer in 2009-2010.

We included 14,596 patients. Median times from bronchoscopy to :

1. neo-adjuvant chemotherapy and to surgery in patients with surgical pathway were **34 days** (Q25=22; Q75=47) and **44 days** (Q25=26; Q75=82), respectively,
2. chemotherapy and to radiotherapy in patients with a non-surgical pathway, were **33 days** (Q25=22; Q75=49) and **88 days** (Q25=46; Q75=162) respectively

Time to first treatment was longer in most northern French regions and in overseas districts and shorter in southern and eastern regions.

Kudjawu YC et al Lung Cancer. 2016


Essai clinique
Critères d' éligibilité
Texte libre, XML, CDISC-ODM
TRADUCTION
Critères d' éligibilité standardisés
Choix d'une syntaxe :
Ad hoc expression, Arden syntax, SQL, GELLO….
Choix d' une terminologie :
UMLS, LOINC, SNOMED CT, READ codes…
Traduction automatique des critères en requêtes
Mapping
Traduction automatique Utilisation d'outils NLP
EDC
Information structurée :
- Données démographiques
- Résultats de biologie
- Codes diagnostiques…
Information non structurée :
- Comptes rendus d' hospitalisation
- Résultats d' anapath
- Résultats de radiologie…
Identification de patients via une plateforme de recrutement

En conclusion

- Du sang, des larmes et du technico-réglementaire
- Mais un énorme potentiel !
  - Taille de l'AP-HP
  - Variétés des données
  - Nombre de chercheurs
  - Données textuelles et génomiques

Ont contribué à cette présentation:

- G Chatellier et A Burgun, Y Girardeau, AS Jannot, S Katsahian, A Neuraz, E Zapletal, (HEGP, Département d'Informatique, Statistiques et Santé Publique)
- E Lenain et J Djadi-Prat, URC HUPO
- P Aegerter et N Bendersky URC Ambroise Paré
- Y Kudjawu, D Eilstein (InVS)
- P Rossignol (CIC-P Nancy)
- C Daniel (AP-HP Equipe Données - WIND - DSI)

- Et tous ceux sans qui nous n'aurions pas de données (Médecins, TECs, TIMS, secrétaires, ingénieurs…) !

Identification des patients pour un essai clinique

- Traduction des critères inclusion
  - Normalisation des critères d'éligibilité :
    - supprimer les critères redondants (inclusion/exclusion)
    - transformation des critères complexes en plusieurs (parfois un grand nombre !) de critères simples
    - reformulation des critères
    - suppression des critères ambigües
    - individualisation des concepts médicaux au sein des critères
- Mapping des concepts médicaux identifiés vers les terminologies utilisées dans l'EDC
- Création des requêtes