Stratégies d'analyse de données en grande dimension issues d'un entrepôt de données cliniques

Anne-Sophie Jannot

annesophie.jannot@aphp.fr

Journées B&S, SFdS – 28 novembre 2016

Plan de l’exposé

- Pourquoi développer des entrepôts de données?
- Bilan de l’expérience à l’HEGP
- Données vie réelle et de grande dimension: quelles difficultés méthodologiques?

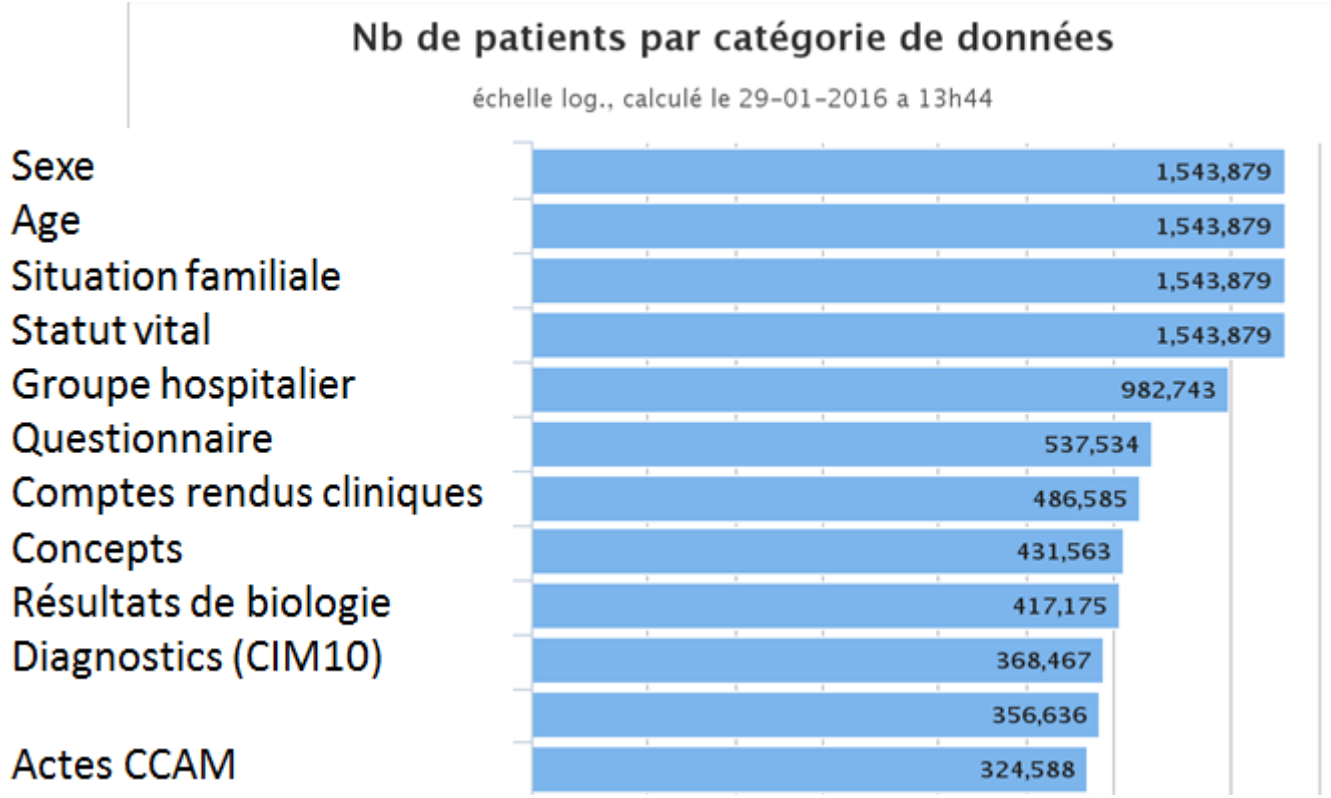
Pourquoi développer des entrepôts de données?

Le concept de médecine personnalisée

- Médecine personnalisée: pouvoir offrir des stratégies de traitement personnalisée en fonction des caractéristiques fines des patients afin d'optimiser le devenir des patients avec la meilleure balance bénéfice/coût/risque.

- Comment estimer l'effet des traitements dans des sous-groupes de patients?
  - Essais cliniques: long, coûteux
  - Réutilisation des données de soin: rapide, peu coûteux mais nécessite:
    - **De trouver des solutions pour stocker les données et pouvoir les réutiliser facilement**
    - D'avoir une taille d'échantillon suffisante
    - De prendre en compte les biais liés à la réutilisation de données en vie réelle

Réutilisation des données du soin: les verrous (1)
Quantité de données

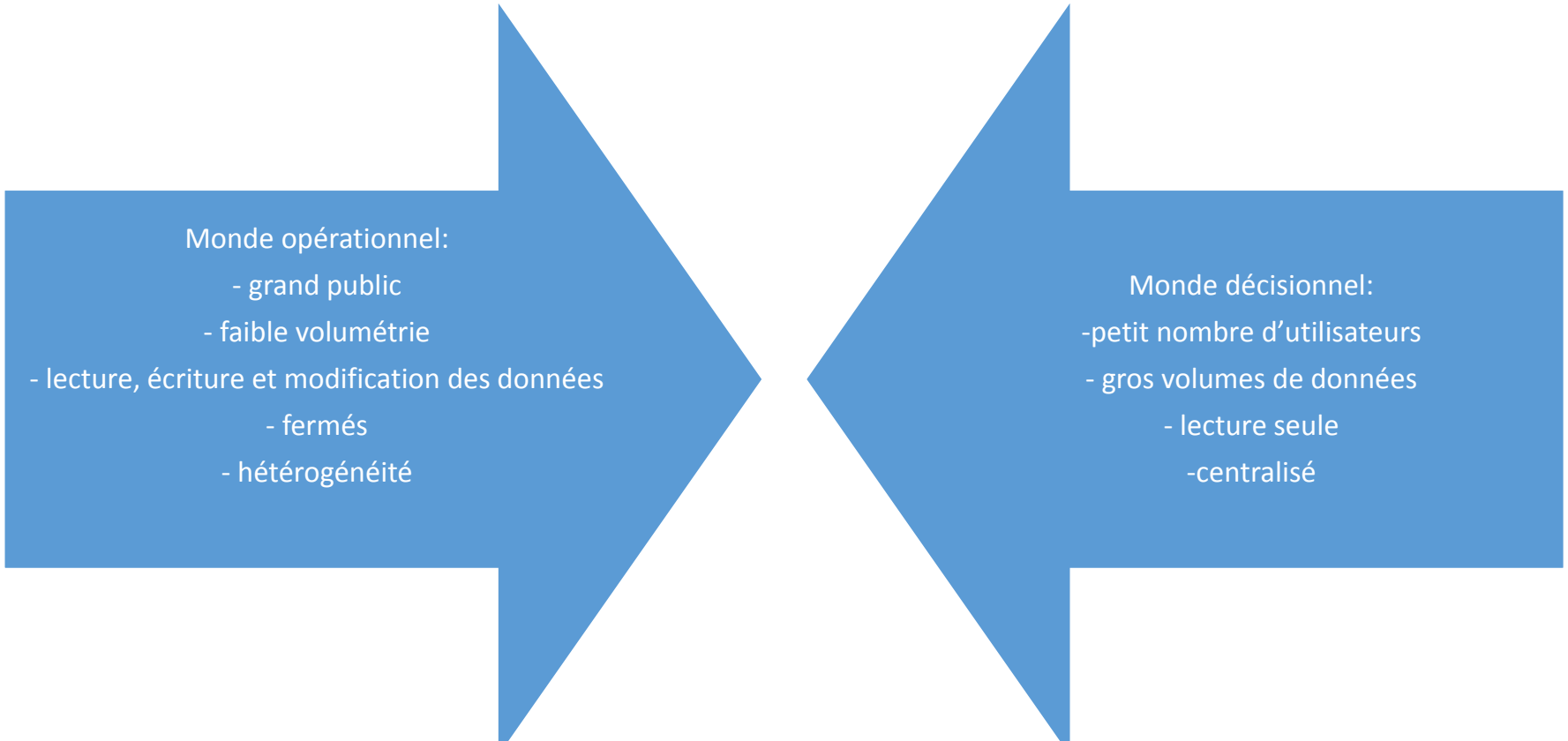


Réutilisation des données du soin: les verrous (2) Données hétérogènes

- Nombreux logiciels métier interconnectés de façon complexe: laboratoire, anatomopathologie, prescription, dispensation, …
- Données de nature différente: textes, valeurs numériques, champs pré-remplis….
- Plusieurs centaines de tables
- Requêtes à partir de plusieurs sources de données très complexes

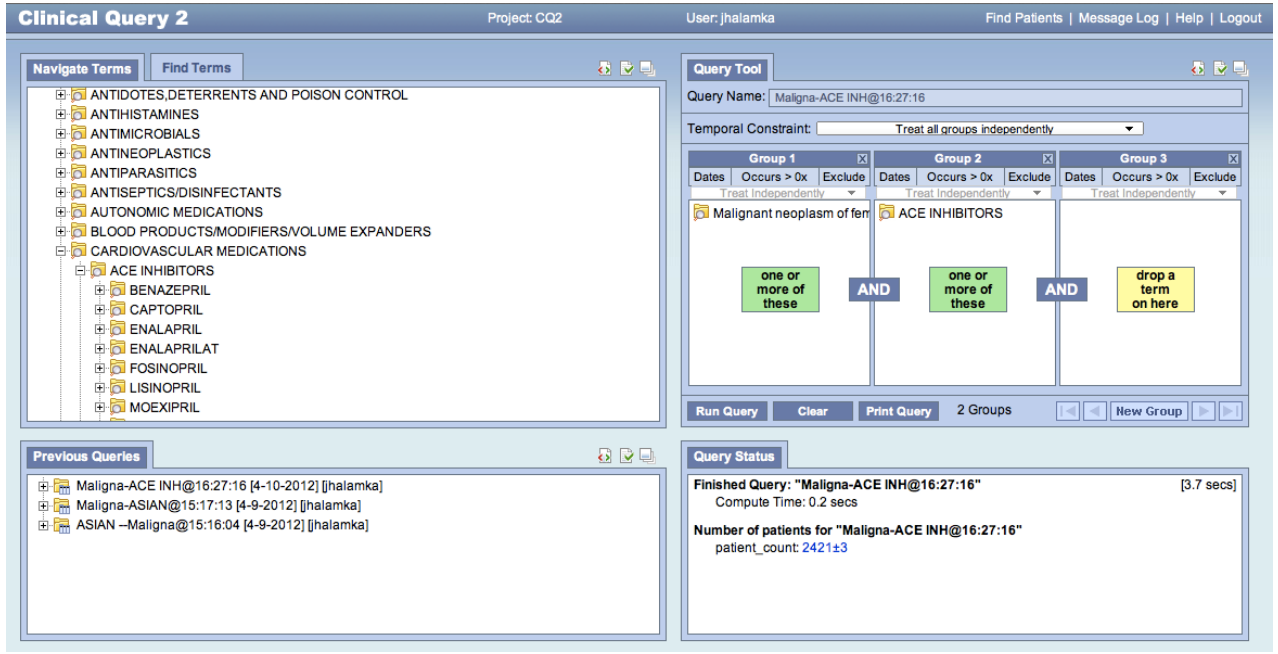
La solution: un entrepôt de données Mais qu'est-ce que c'est?

- Concept issu de la « business intelligence »



Un modèle d'entrepôt: la plateforme i2b2

- Initiative développée par l'Université d'Harvard permettant de:
  - Définir un format commun de données hospitalières
  - Proposer des outils de visualisation et d'analyse des données cliniques
- Nombreuses institutions ont mis en place cette plateforme



Pour plus d'informations:
https://www.youtube.com/watch?v=BrzW2FxwO1M

Intégration d'un entrepôt au sein d'un système d'information: plusieurs objectifs

- **Recherche**
  - Epidémiologie: étude cas/témoins, recherche de signaux / vigilance
  - Recherche sur les services en santé : évaluation in silico « en vie réelle »
  - Recrutement des patients dans les essais cliniques
- **Pilotage de l'hôpital**
  - Recrutement des patients
  - Test « in silico » de changements organisationnels
  - Optimisation du codage
- **Amélioration des soins**
  - Recherche de patients similaires
  - Evaluation des pratiques et mise en place d'outils d'aide à la décision
  - Elaboration de parcours type de patients

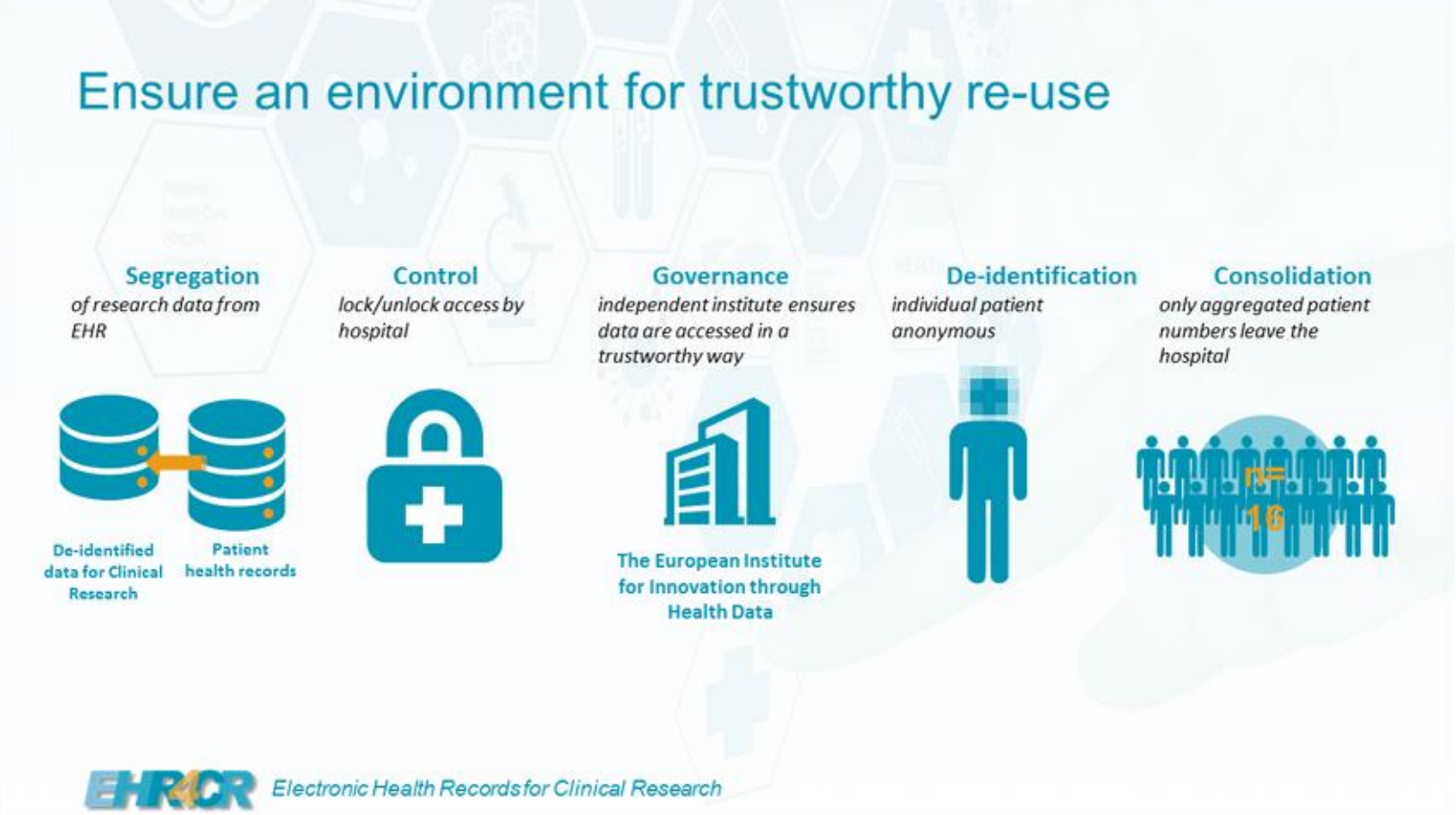
Des dossiers patients à la recherche : Unicancer à l'heure des big data

Le prototype du projet ConSoRe vient tout juste d'être finalisé et l'assemblée générale d'Unicancer a validé son déploiement : les centres de lutte contre le cancer (CLCC) sont en train de se doter d'un puissant outil de recherche sémantique permettant d'interroger l'ensemble de leurs bases de données, quels que soient leur contenu et leur structure.

- Avant: identification des patients ayant le profil requis pour un essai clinique passait par une revue manuelle particulièrement laborieuse des dossiers patients comprenant des millions de documents pour chaque centre
- ConSoRe: outil de recherche permettant, pour une étude, de constituer rapidement des cohortes de patients réunissant l'ensemble des critères requis (sexe, type de tumeur, traitement reçu...).

→ Accélération de la mise en place des essais cliniques

Réutilisation pour la recherche clinique



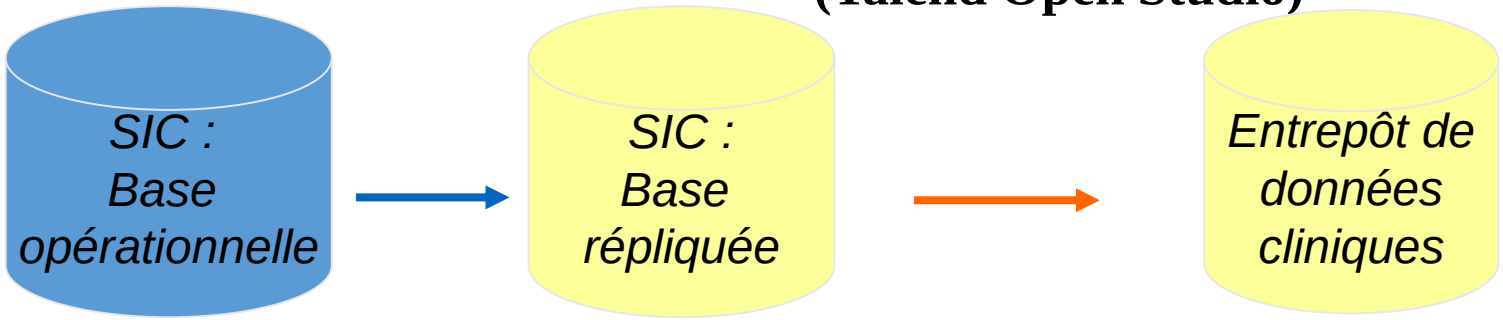
http://www.ehr4cr.eu

Pour en savoir plus: https://www.youtube.com/watch?v=Wcsl064F2pk

Bilan de l’expérience à l’HEGP

Intégration de l'entrepôt au SIH de l'HEGP (2009-)

Suite Extract / Transform / Load
(Talend Open Studio)

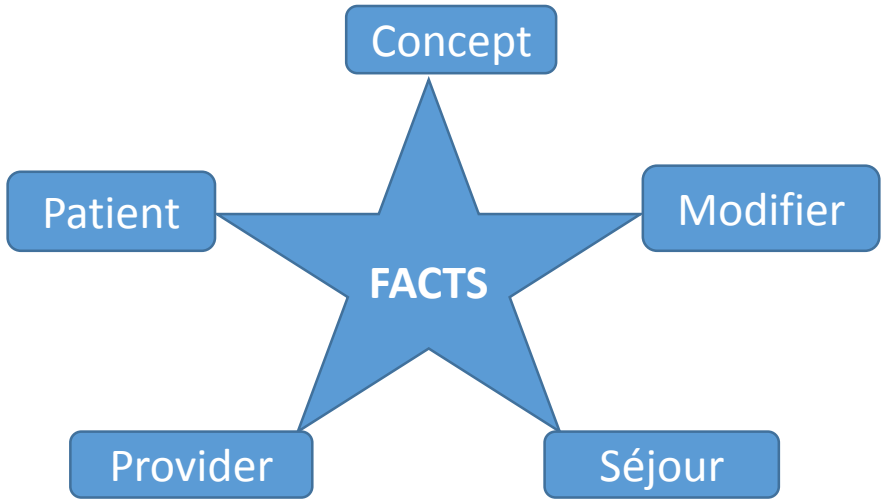


EDASYS EDASYS

i2b2
Informatics for Integrating Biology & the Bedside

Dossier Patient informatisé

- Prescription
- Comptes rendus
- Biologie
- …



Processus de fonctionnement de l'entrepôt: projet de recherche

- Complétion du document pour le comité d'éthique: définition des variables à extraire
- Comité d'éthique: vérification du niveau d'accès aux données
- Extraction des patients et de leurs caractéristiques associées (sql)
- Visualisation des informations non structurées, analyse des données

Répartition des projets menés depuis la mise à disposition de l'entrepôt

Année	Nombre de projets	Nombre de départements	Epidémiologie clinique	Recherche sur les services en santé	Recherche clinique
2011	13	4	8	5	0
2012	6	4	3	3	0
2013	13	6	8	4	1
2014	22	11	15	5	2
2015	22	15	9	13	0
Total (%)	76 (100%)	18 (75%)	38 (50%)	26 (34%)	12 (16%)

Données vie réelle et de grande dimension: quelles difficultés méthodologiques?

1ère partie: étude épidémiologique – analyse du lien hyponatrémie / mortalité

2ème partie: évaluation des pratiques en santé – l'injection de produit de contraste induit-elle une insuffisance rénale aigue?

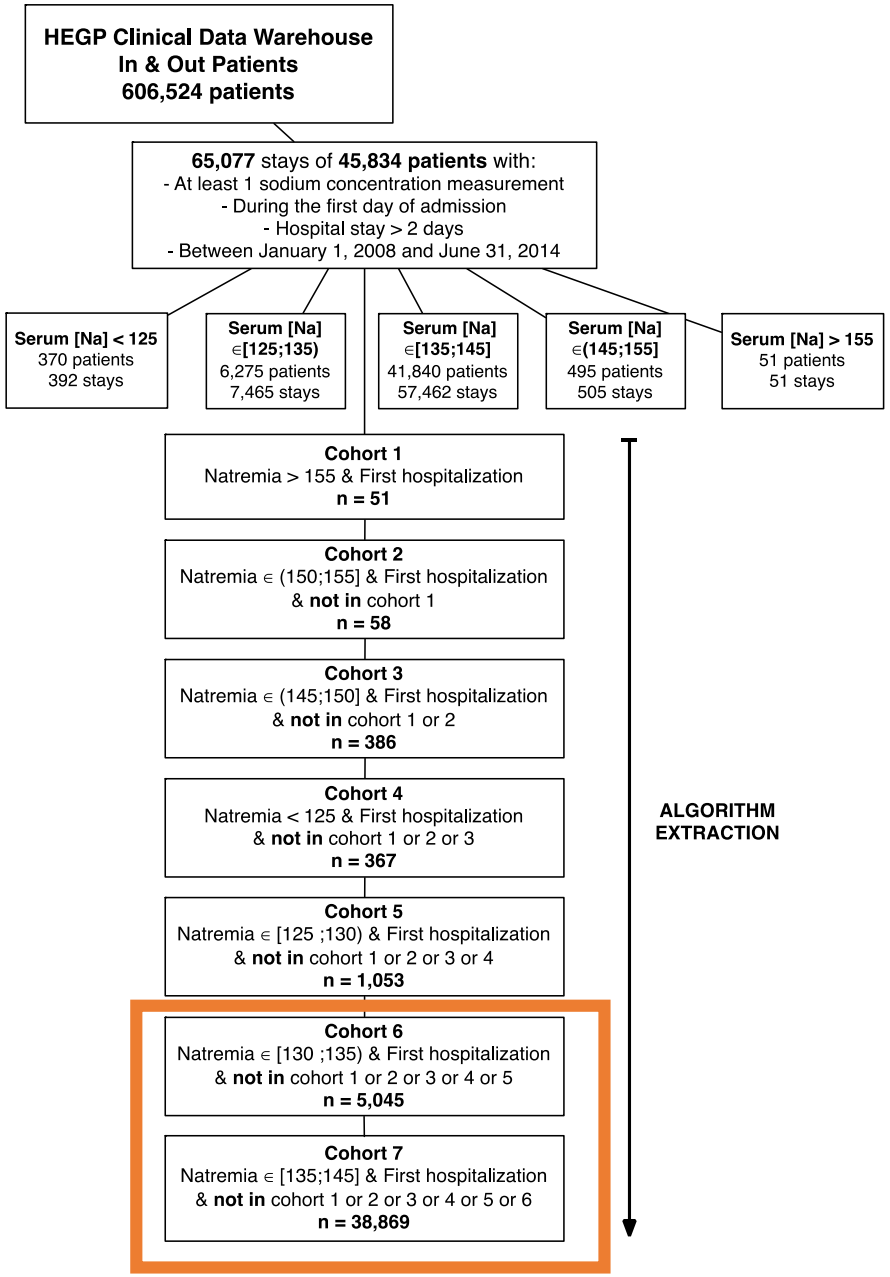
Exemple d'étude épidémiologique: Natrémie et mortalité / stratégie PheWAS

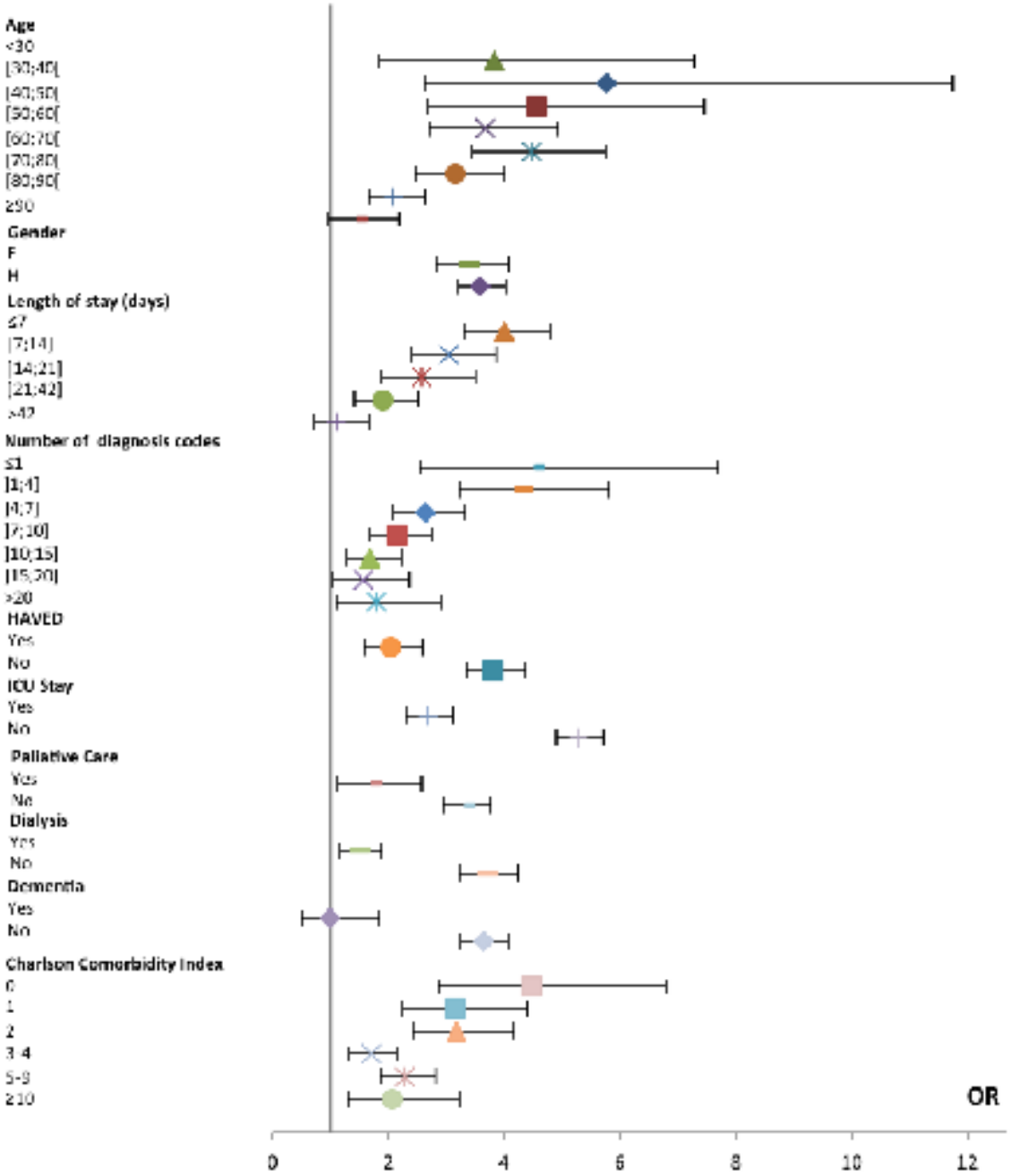
- Question de recherche: existe-t-il un lien entre hyponatrémie « borderline » (entre 130 et 135 mmol/l) et mortalité? Quelle est la force de l'association?

- Etude pronostique sur des données historiques (avantage de l'entrepôt: plus de 15 ans d'historique de données)

- Prise en compte des facteurs de confusion de façon exhaustive possible:
  - D'après la littérature: très nombreux
  - D'après une stratégie de type PheWAS (fouille de données): recherche exhaustive de tous les diagnostics pouvant constituer des facteurs de confusion

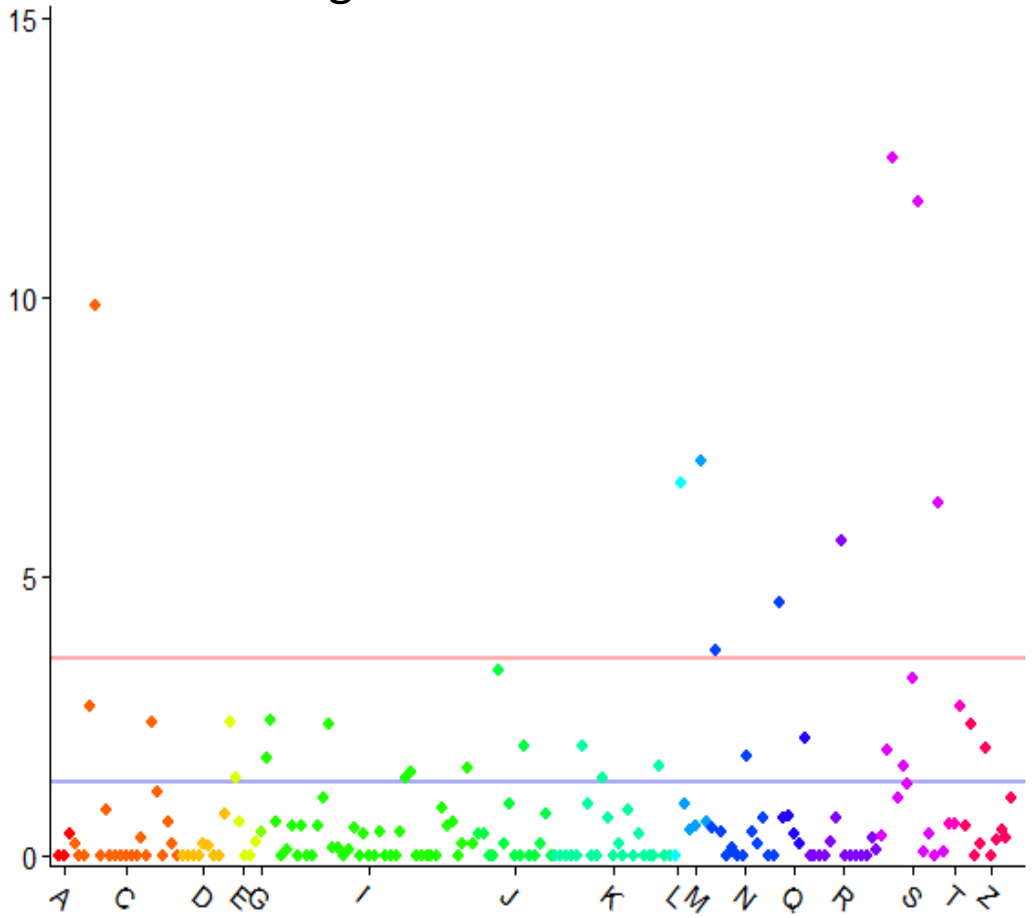
Association between Borderline Dysnatremia and Mortality Insight Into a New Data Mining Approach . Yannick Girardeau, Anne-Sophie Jannot, Gilles Chatellier, Olivier Saint-Jean


HEGP Clinical Data Warehouse
In & Out Patients
606,524 patients
65,077 stays of 45,834 patients with:
- At least 1 sodium concentration measurement
- During the first day of admission
- Hospital stay > 2 days
- Between January 1, 2008 and June 31, 2014
Serum [Na] < 125
370 patients
392 stays
Serum [Na] ∈[125;135)
6,275 patients
7,465 stays
Serum [Na] ∈[135;145]
41,840 patients
57,462 stays
Serum [Na] ∈(145;155]
495 patients
505 stays
Serum [Na] > 155
51 patients
51 stays
Cohort 1
Natremia > 155 & First hospitalization
n = 51
Cohort 2
Natremia ∈ (150;155] & First hospitalization
& not in cohort 1
n = 58
Cohort 3
Natremia ∈ (145;150] & First hospitalization
& not in cohort 1 or 2
n = 386
Cohort 4
Natremia < 125 & First hospitalization
& not in cohort 1 or 2 or 3
n = 367
Cohort 5
Natremia ∈ [125 ;130) & First hospitalization
& not in cohort 1 or 2 or 3 or 4
n = 1,053
Cohort 6
Natremia ∈ [130 ;135) & First hospitalization
& not in cohort 1 or 2 or 3 or 4 or 5
n = 5,045
Cohort 7
Natremia ∈ [135;145] & First hospitalization
& not in cohort 1 or 2 or 3 or 4 or 5 or 6
n = 38,869
ALGORITHM
EXTRACTION

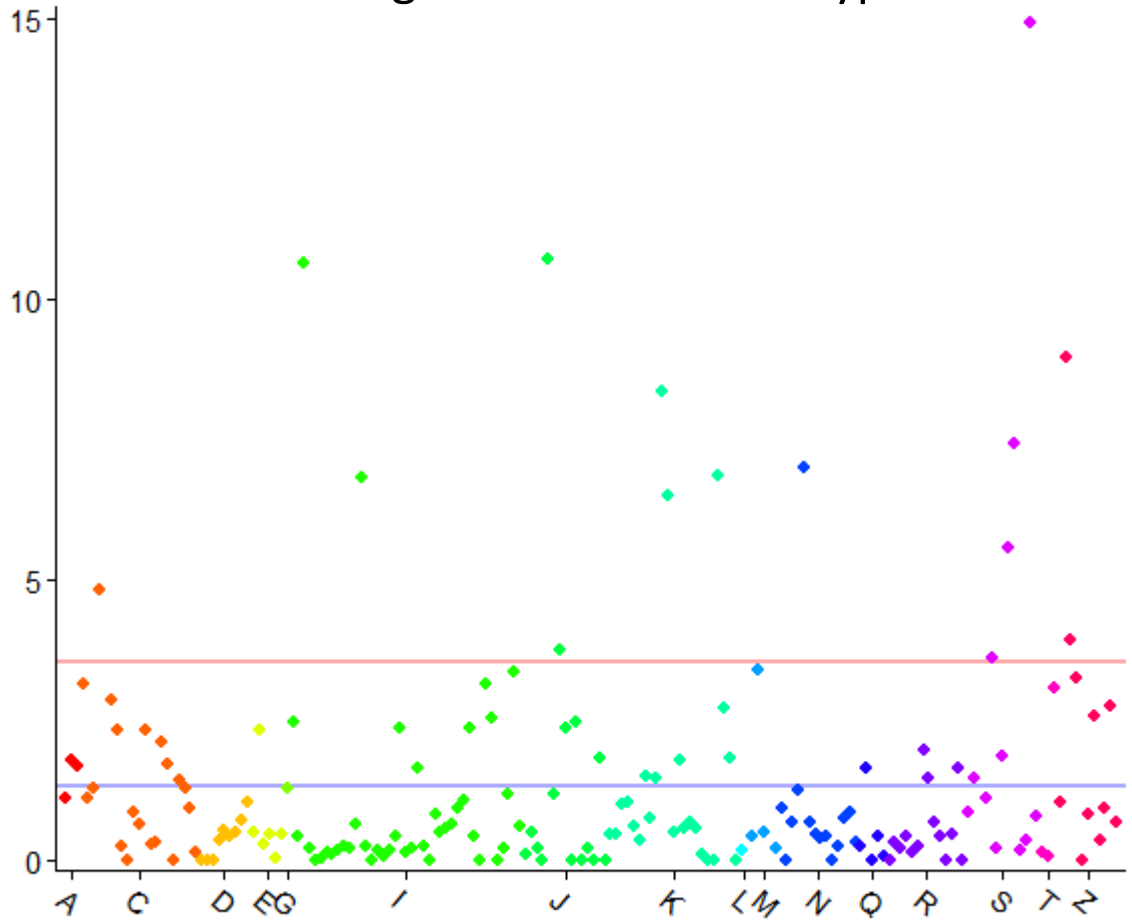

Age
<30
[30;40[
[40;50[
[50;60[
[60;70[
[70;80[
[80;90[
≥90
Gender
F
H
Length of stay (days)
≤7
[7;14]
[14;21]
[21;42]
>42
Number of diagnosis codes
≤1
[1;4]
[4;7]
[7;10]
[10;15]
[15;20]
>20
HAVED
Yes
No
ICU Stay
Yes
No
Palliative Care
Yes
No
Dialysis
Yes
No
Dementia
Yes
No
Charlson Comorbidity Index
0
1
2
3-4
5-9
≥10
OR
0
2
4
6
8
10
12

Stratégie « PheWAS » et recherche de facteurs de confusion

Association diagnostics CIM-10 et mortalité



Association diagnostics CIM-10 et hyponatrémie limite



ICD-10 codes	Mortality	P Value	Hyponatremmia	P Value
A41 Other sepsis	4.7 (3.1; 6.88)	< 0.001	5.83 (4.4; 7.69)	< 0.001
I20 Angina pectoris	0.23 (0.13; 0.37)	< 0.001	0.67 (0.55; 0.79)	< 0.001
I25 Chronic ischemic heart disease	0.47 (0.34; 0.64)	< 0.001	0.53 (0.44; 0.62)	< 0.001
I48 Atrial fibrillation and flutter	0.15 (0.07; 0.3)	< 0.001	0.62 (0.5; 0.76)	< 0.001
I71 Aortic aneurysm and dissection	2.37 (1.9; 2.92)	< 0.001	0.66 (0.53; 0.8)	< 0.001
J15 Bacterial pneumonia, not elsewhere classified	2.9 (1.94; 4.16)	< 0.001	2.75 (2.13; 3.51)	< 0.001
J80 Acute respiratory distress syndrome	30.64 (23.27; 40.36)	< 0.001	3.1 (2.28; 4.17)	< 0.001
J96 Respiratory failure, not elsewhere classified	6.14 (5.19; 7.22)	< 0.001	2.29 (1.98; 2.63)	< 0.001
K65 Peritonitis	6.53 (4.46; 9.3)	< 0.001	3.19 (2.32; 4.33)	< 0.001
R07 Pain in throat and chest	0.1 (0.02; 0.3)	< 0.001	0.35 (0.23; 0.5)	< 0.001
R57 Shock, not elsewhere classified	18.73 (15.81; 22.16)	< 0.001	3.56 (3.01; 4.21)	< 0.001
Z48 Encounter for other postprocedural aftercare	0.63 (0.51; 0.77)	< 0.001	0.62 (0.55; 0.7)	< 0.001
Z51 Encounter for other aftercare	2.46 (1.97; 3.04)	< 0.001	1.95 (1.68; 2.25)	< 0.001

Résultats finaux et conclusion

Association Between Borderline Hyponatremia and Mortality		
	OR (IC95%)	P
Classical Model	1.98 (1.73;2.68)	<.001
Phewas Model	2.59 (2.28;2.94)	<.001
Final Model	1.57 (1.35;1.81)	<.001

Stratégie « PHeWAS » ne remplace pas l'expertise, mais la complète

Données vie réelle et de grande dimension: quelles difficultés méthodologiques?

1ère partie: étude épidémiologique – analyse du lien hyponatrémie / mortalité

2ème partie: évaluation des pratiques en santé – l'injection de produit de contraste induit-elle une insuffisance rénale aigue?

Intravenous Contrast Material–induced Nephropathy: Causal or Coincident Phenomenon?[1]

Purpose: To determine the causal association and effect of intravenous iodinated contrast material exposure on the incidence of acute kidney injury (AKI), also known as contrast material–induced nephropathy (CIN).

Materials and Methods: This retrospective study was approved by an institutional review board and was HIPAA compliant. Informed consent was waived. All contrast material–enhanced (contrast group) and unenhanced (noncontrast group) abdominal, pelvic, and thoracic CT scans from 2000 to 2010 were identified at a single facility. Scan recipients were sorted into low- (<1.5 mg/dL), medium- (1.5–2.0 mg/dL), and high-risk (>2.0 mg/dL) subgroups of presumed risk for CIN by using baseline serum creatinine (SCr) level. The incidence of AKI (SCr ≥ 0.5 mg/dL above baseline) was compared between contrast and noncontrast groups after propensity score adjustment by stratification, 1:1 matching, inverse weighting, and weighting by the odds methods to reduce intergroup selection bias. Counterfactual analysis was used to evaluate the causal relation between contrast material exposure and AKI by evaluating patients who underwent contrast-enhanced and unenhanced CT scans during the study period with the McNemar test.

McDonald, R. J., McDonald, J. S., Bida, J. P., Carter, R. E., Fleming, C. J., Misra, S., ... & Kallmes, D. F. (2013). Intravenous contrast material–induced nephropathy: causal or coincident phenomenon?. Radiology, 267(1), 106-118.

Question de santé publique +++

Le contexte de l’étude

Traitement = exposition = avoir une injection de produit de contraste

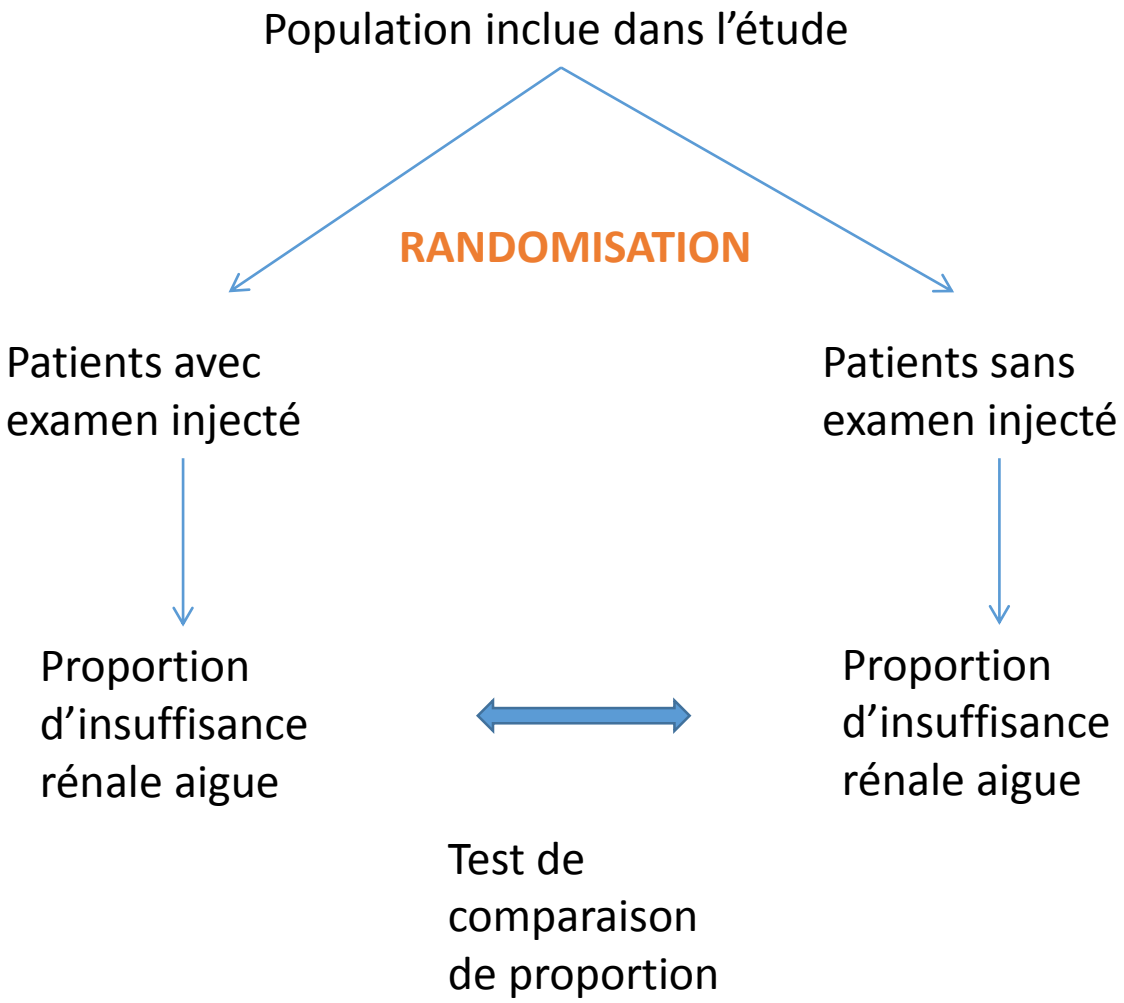
Le contexte de l’étude

Effet secondaire = Maladie = insuffisance rénale aigue = élévation de la créatinine = baisse du débit de filtration glomérulaire

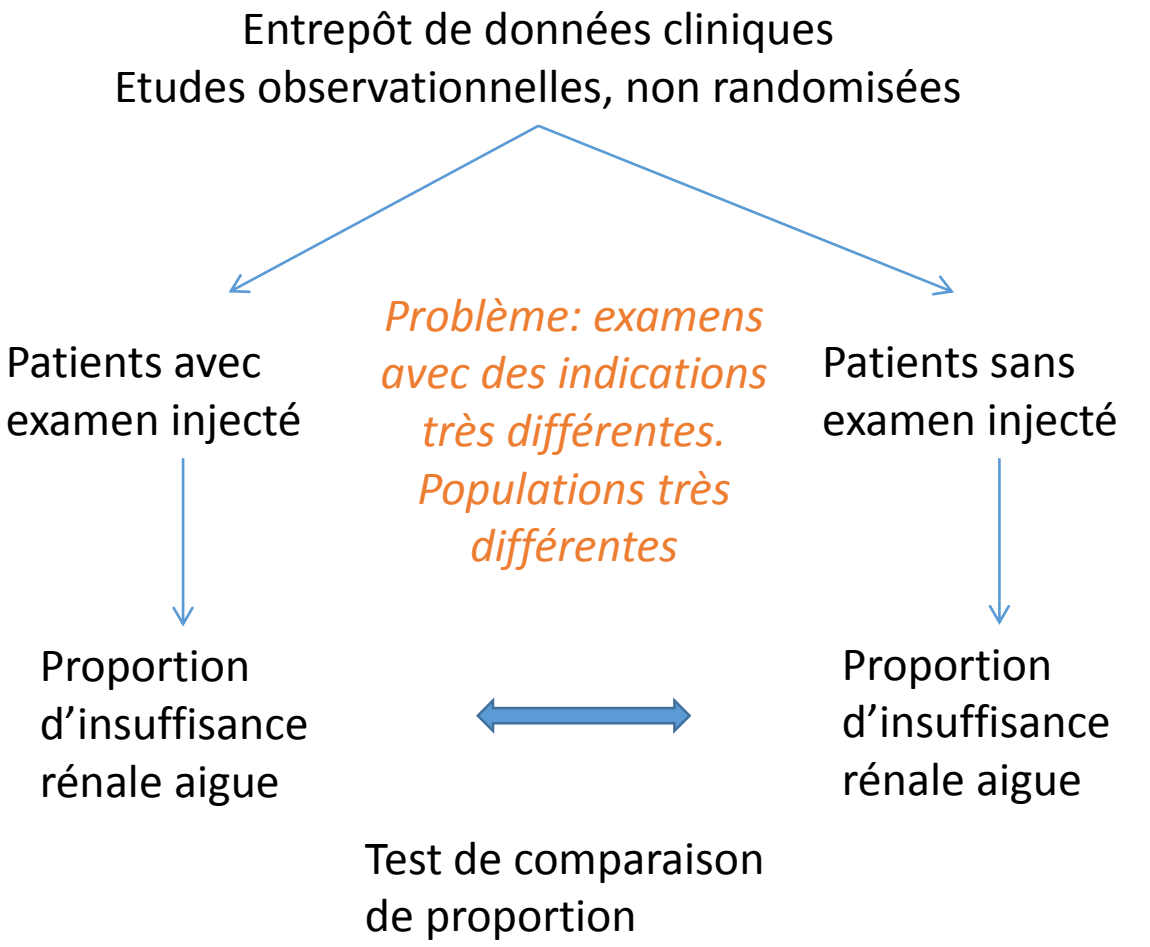
Le contexte de l’étude

- Question de recherche: y a-t-il un lien entre injection de produit de contraste et insuffisance rénale aigue?
- Comparaison de deux groupes:
  - Ceux ayant eu un scan injecté
  - Ceux n’ayant pas eu de scan injecté
- Critère de jugement (effet traitement): avoir une insuffisance rénale aigue

Le monde idéal: l'essai interventionnel



Le monde réel: randomisation impossible



Nécessité de prendre en compte les facteurs de confusion

Comment estimer l'effet d'un traitement à partir de données observationnelles?

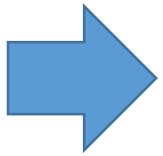
- Comment estimer l'effet « traitement »?
- Sur le plan mathématique:
  - Soit Y, l'effet traitement étudié (ex: effet secondaire, efficacité)
  - Soit un individu i
  - Soit Y1i, le résultat s'il reçoit le traitement
  - Soit Y0i, le résultat s'il ne reçoit pas le traitement
- L'effet traitement est estimé par la moyenne de la différence sur l'ensemble des individus
- Impossible avec des données observationnelles

Les méthodes « classiques » pour prendre en compte les facteurs de confusion

- ***Ajustement:***
  - ***en pratique***: analyse multivariée.
- ***Stratification:***
  - ***en pratique****:* chaque strate est un niveau du facteur de confusion, soit 2^k strates pour k facteurs de confusion dichotomiques.
- ***Appariement:***
  - ***en pratique***: pour chaque cas (exemple :un malade), on associe un ou plusieurs témoins qui lui sont similaires pour un ou plusieurs facteurs (exemple: âge, sexe, niveau socio-économique).

En pratique: impossible!

- Entrepôt de données: plusieurs centaines de variables par individu
  - Appariement impossible
  - Stratification: problème des très faibles effectifs par strate
  - Ajustement: violation des hypothèses des modèles multivariés



Méthode de choix: score de propension?

Le score de propension

Deux étapes:

- Estimation de la **probabilité d'appartenance au groupe traité et non traité en fonction des caractéristiques de l'individu** = score de propension
  - Permet de « résumer » l'information sur les facteurs de confusion en une seule valeur
- **Appariement** des patients qui ont à peu près le même score de propension

A chaque étape, difficultés méthodologiques pouvant induire un biais important

Comment estimer le score de propension?

- Toute méthode qui permet d'estimer une probabilité d'appartenance à un groupe en fonction de variables
- Méthode classique : la régression logistique multivariée:
  - Variable à expliquer: le fait d'être traité (variable binaire)
  - Variables explicatives: les facteurs de confusion
  - Limites de la régression logistique:
    - Combinaison linéaire
    - Problème des convergence lorsque les facteurs de confusion sont très corrélés
    - Nombre de variables pouvant être pris en compte limité
- Méthodes « modernes »: régression pénalisée, arbre de décision…

Limites: imprécision dans l'estimation du score de précision liée à la fois aux facteurs de confusion non disponibles et au modèle choisie

Procédure d'appariement

- Méthode la plus utilisée : « nearest neighbour matching »: chaque individu i traité est apparié avec le patient j du groupe contrôle ayant le sore de propension le plus proche

$$|p(X_i) - p(X_j)| \leq \epsilon SD(p(X)).$$

$SD(p(X))$ = déviation standard du score de propension dans la population

ϵ = caliper

Choix du caliper: compromis réduction du biais / puissance de l'étude
Recommandation: caliper = 0.2 mais…

Une petite étude de simulation

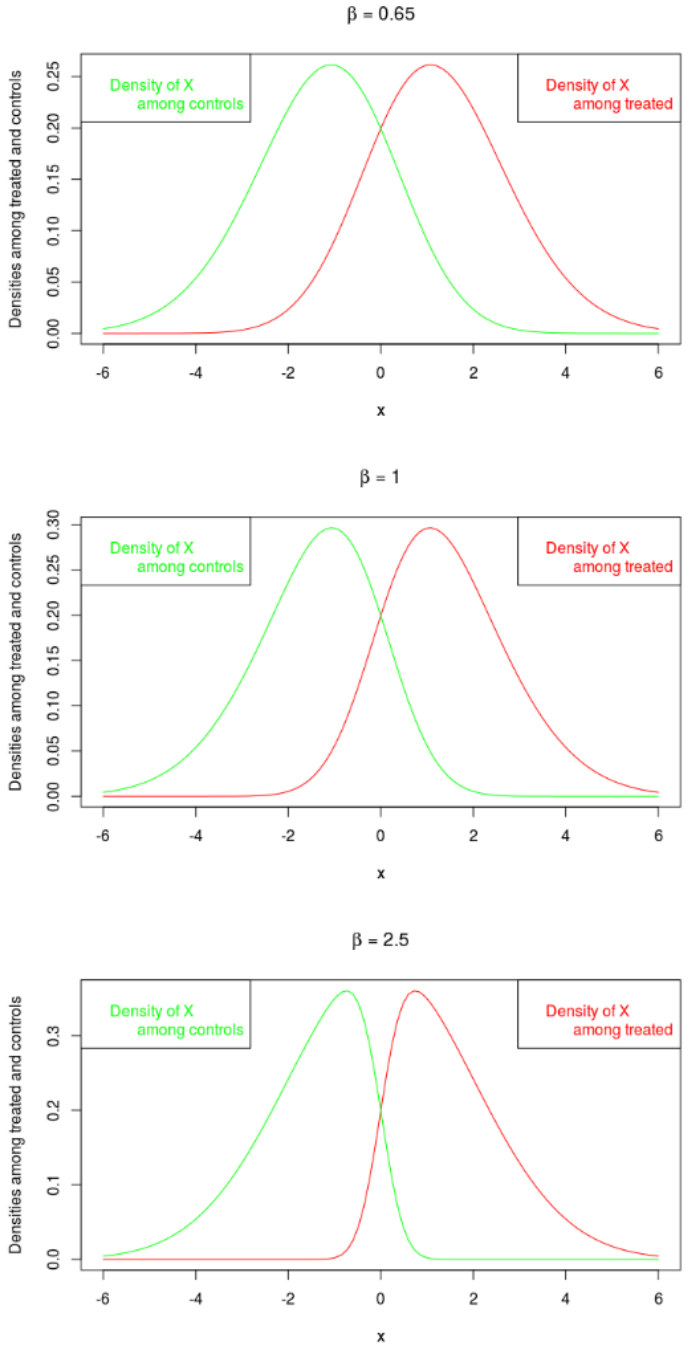
$$X \sim \mathcal{N}(0, \sigma_x^2)$$

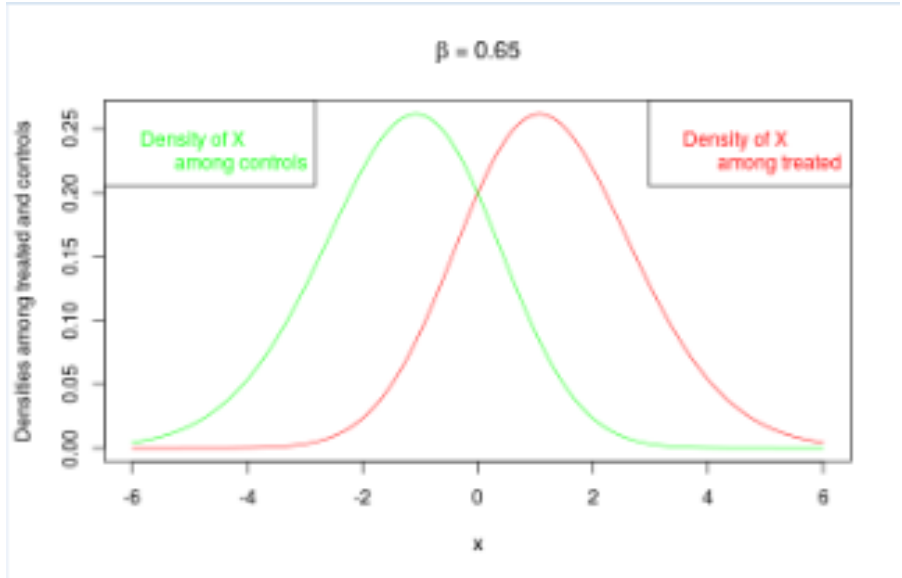
$$\mathbb{P}(T = 1 \mid X = x) = \Phi(\beta x)$$

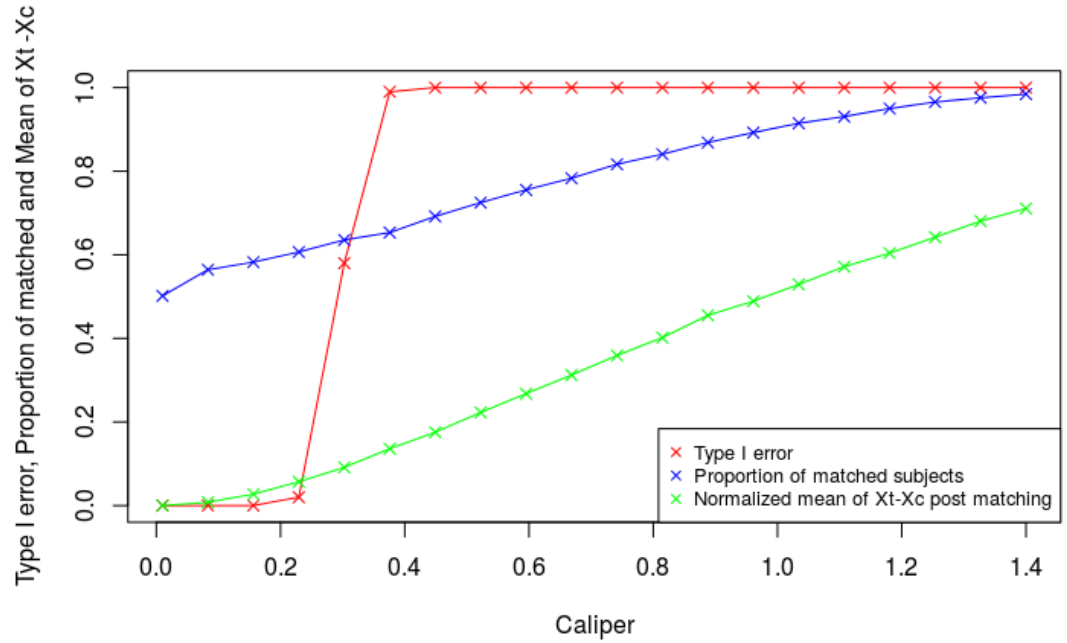
$$Y = a + b\,X + c\,T + \eta$$

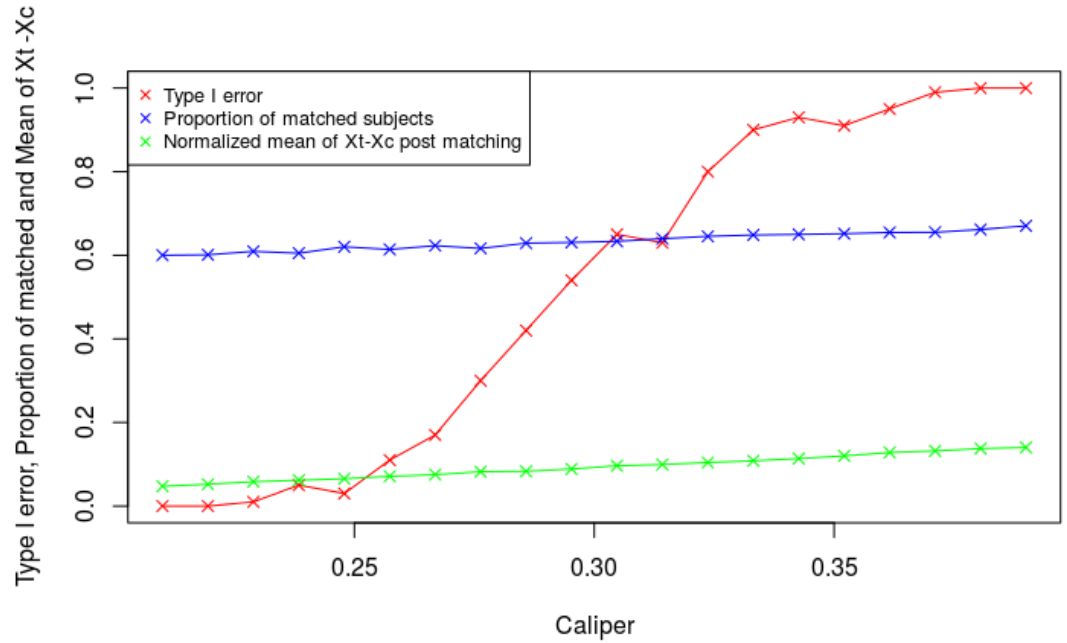
$$a = 0,\ b = 1,\ \sigma_\eta = 0.1$$

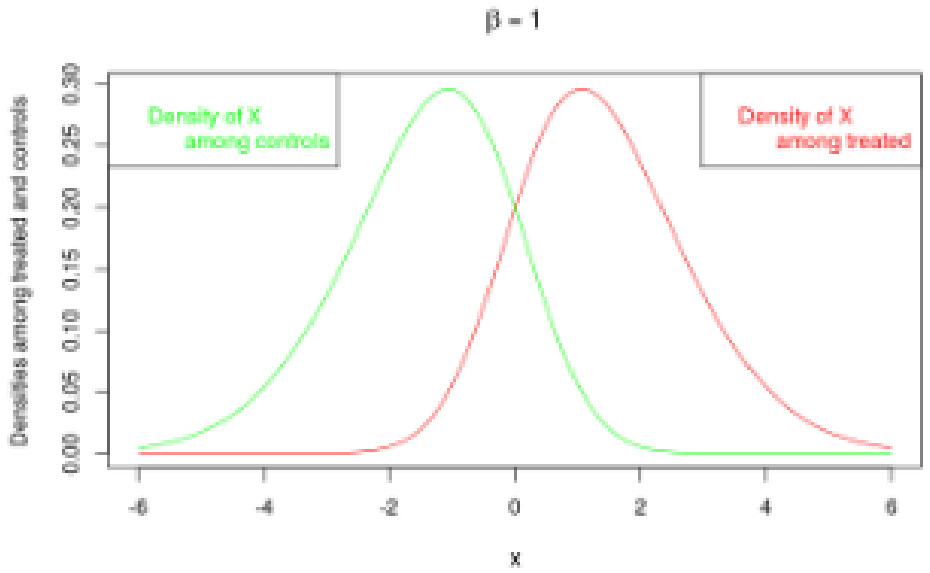
No universal recommendation for caliper choice is possible when using propensity score matching
Emeline Fay, Agathe Guilloux, Anne-Sophie Jannot

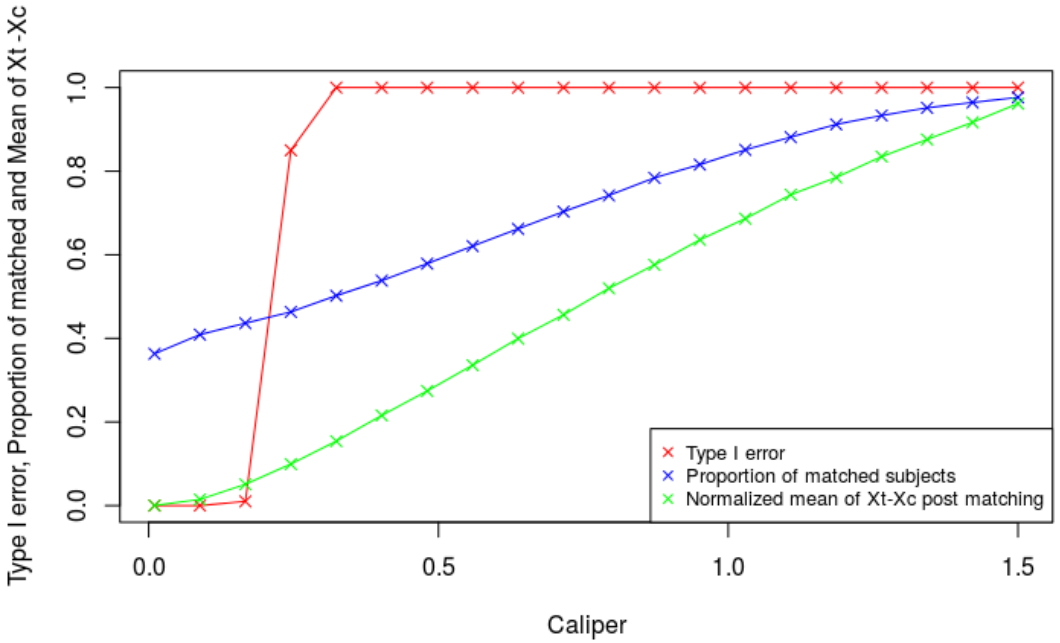


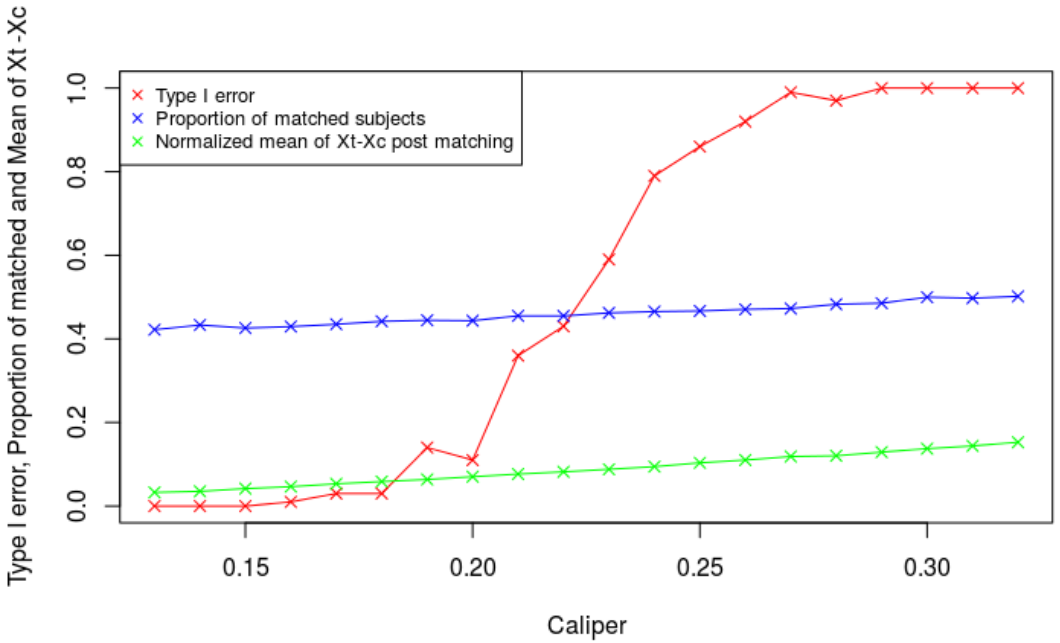

β = 0.65
Density of X
among controls
Density of X
among treated
Densities among treated and controls
0.00
0.05
0.10
0.15
0.20
0.25
-6
-4
-2
0
2
4
6
x

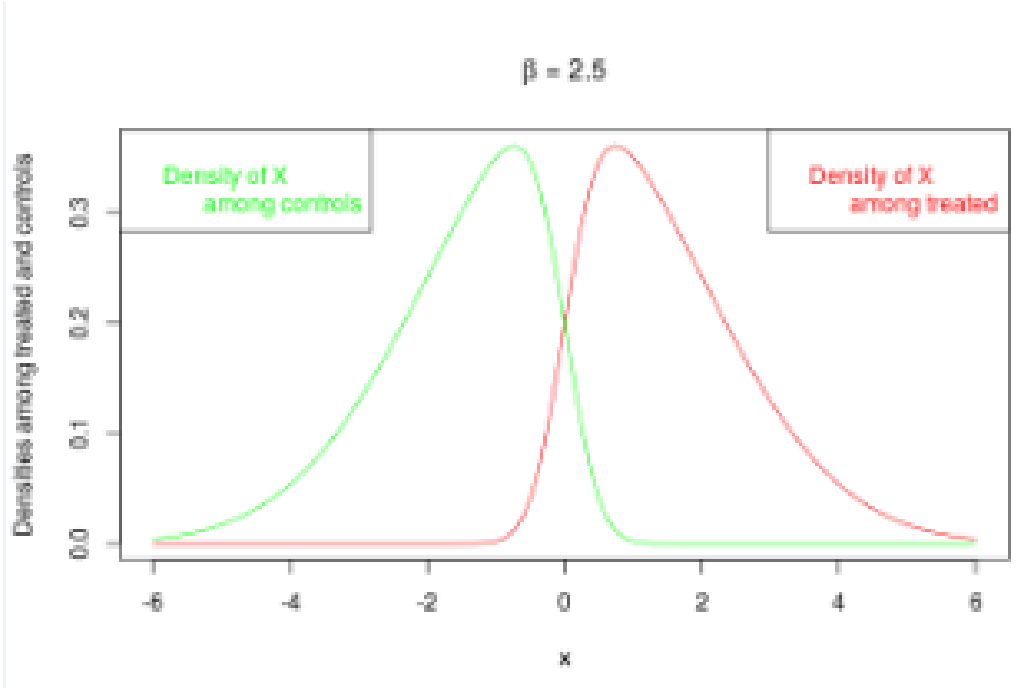

Type I error, Proportion of matched and Mean of Xt -Xc
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Caliper
Type I error
Proportion of matched subjects
Normalized mean of Xt-Xc post matching

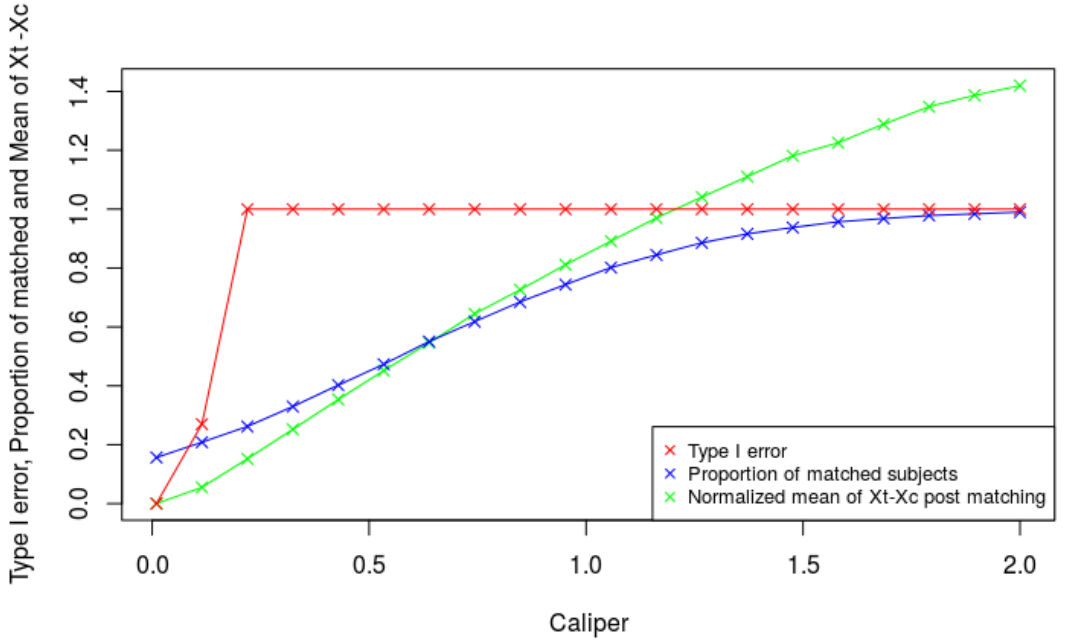

Type I error, Proportion of matched and Mean of Xt -Xc
0.0
0.2
0.4
0.6
0.8
1.0
0.25
0.30
0.35
Caliper
Type I error
Proportion of matched subjects
Normalized mean of Xt-Xc post matching

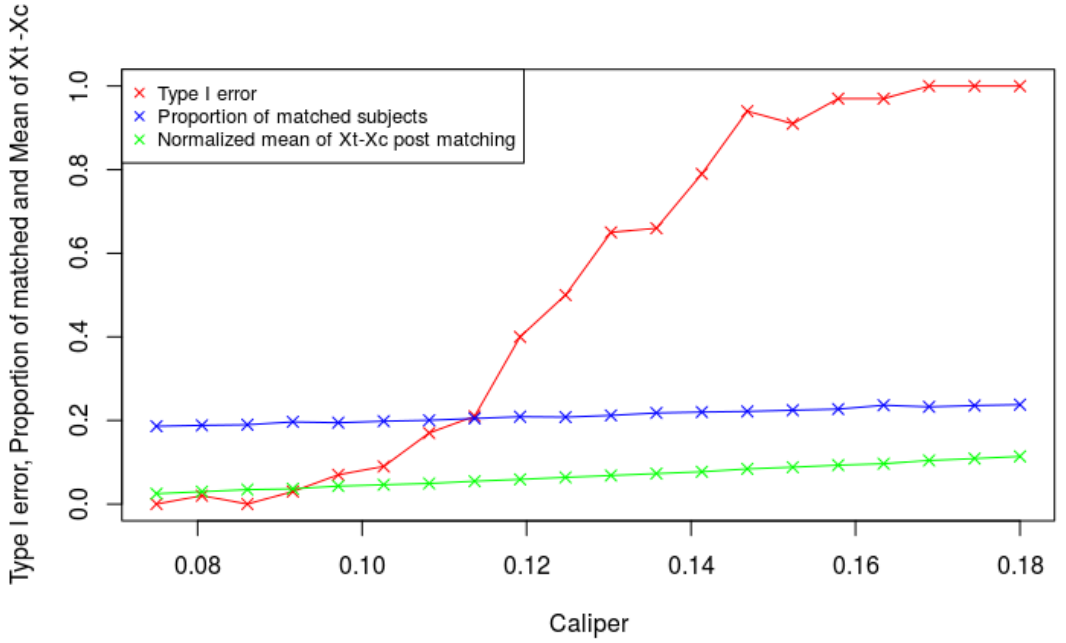

β = 1
Densities among treated and controls
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Density of X
among controls
Density of X
among treated
-6
-4
-2
0
2
4
6
x


Type I error, Proportion of matched and Mean of Xt -Xc
0.0
0.2
0.4
0.6
0.8
1.0
Type I error
Proportion of matched subjects
Normalized mean of Xt-Xc post matching
0.0
0.5
1.0
1.5
Caliper


Type I error, Proportion of matched and Mean of Xt -Xc
0.0
0.2
0.4
0.6
0.8
1.0
Type I error
Proportion of matched subjects
Normalized mean of Xt-Xc post matching
0.15
0.20
0.25
0.30
Caliper







0.2 ne peut être considéré comme un bon choix de caliper quelles que soient les distributions de X dans les deux groupes

Conclusion

- Données d'entrepôt:
  - Très grand nombre de variables disponibles par individus
  - Données hétérogènes

- Stratégies de type fouille de données complètent mais ne remplacent pas l'expertise

- Inférence causale pour l'évaluation des pratiques en santé: le score de propension peut être utilisé mais uniquement pour comparer des stratégies qui s'adressent au même type de patients

Remerciements

- Groupe Biopharmacie et Santé – SFDS

- Anita Burgun, Eric Zapletal, Patrice Degoulet et toute l'équipe entrepôt de l'HEGP

- Equipe 22 – UMRS 1138 « Science de l'information au service de la médecine personnalisée », Centre de Recherche des Cordeliers

Merci pour votre attention

A post-doc position for one to two years is open in Paris and Kyoto to work on "Bridging and modeling differences and similarities of heterogeneous populations in medical cancer care using medical data warehouses" at the French Institute of Health and Medical Research (INSERM) and at the Graduate School of Medicine and Faculty of Medicine Kyoto University.

Contacts: sarah.zohar@inserm.fr , annesophie.jannot@aphp.fr and smorita@kuhp.kyoto-u.ac.jp