Utilisation du score de propension à haute dimension

Recherche de comorbidités liées à la consommation excessive d'alcool

Clément TEISSIER

CEMKA
Maitre de stage : Stéphane BOUEE

Avril - Septembre 2018

CEMKA

- ➔ Bureau d'études spécialisé dans la santé
- ➔ Stage effectué au sein du pôle Biométrie et Analyse de Bases de Données

Contexte

Données : Cohorte **CONSTANCES** appariée au **SNIIRAM**

➔ But : chercher les liens entre la **consommation d'alcool** et certaines **comorbidités.**
➔ Stage mené en lien avec l'unité UMS-011 de l'INSERM : cohortes épidémiologiques

Contexte : la cohorte CONSTANCES

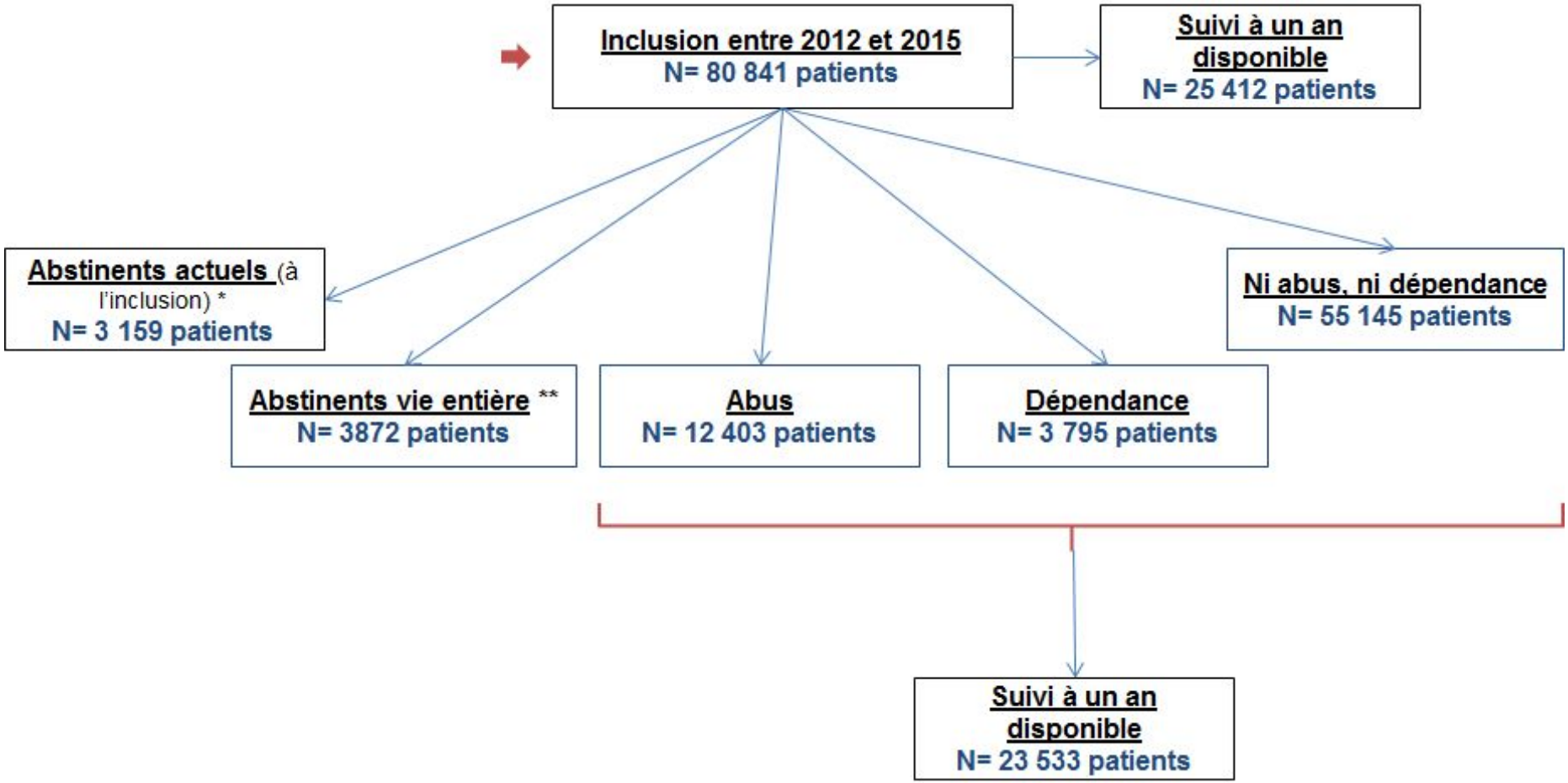
- ➔ Cohorte épidémiologique constituée par l'INSERM
- ➔ 200 000 adultes (18 à 69 ans à l'inclusion) affiliés au régime général de la sécurité sociale
- ➔ Questionnaire et examen médical à l'inclusion, à partir de 2012
- ➔ Questionnaire de suivi tous les ans, examens tous les 5 ans
- ➔ Données diverses
  - ◆ Données socio-démographiques
  - ◆ Antécédents médicaux
  - ◆ Consommations d'alcool, tabac, cannabis
  - ◆ Données biologiques à l'inclusion (taille, poids, IMC, pression artérielle...)
  - ◆ Autres : alimentation, vie sexuelle…

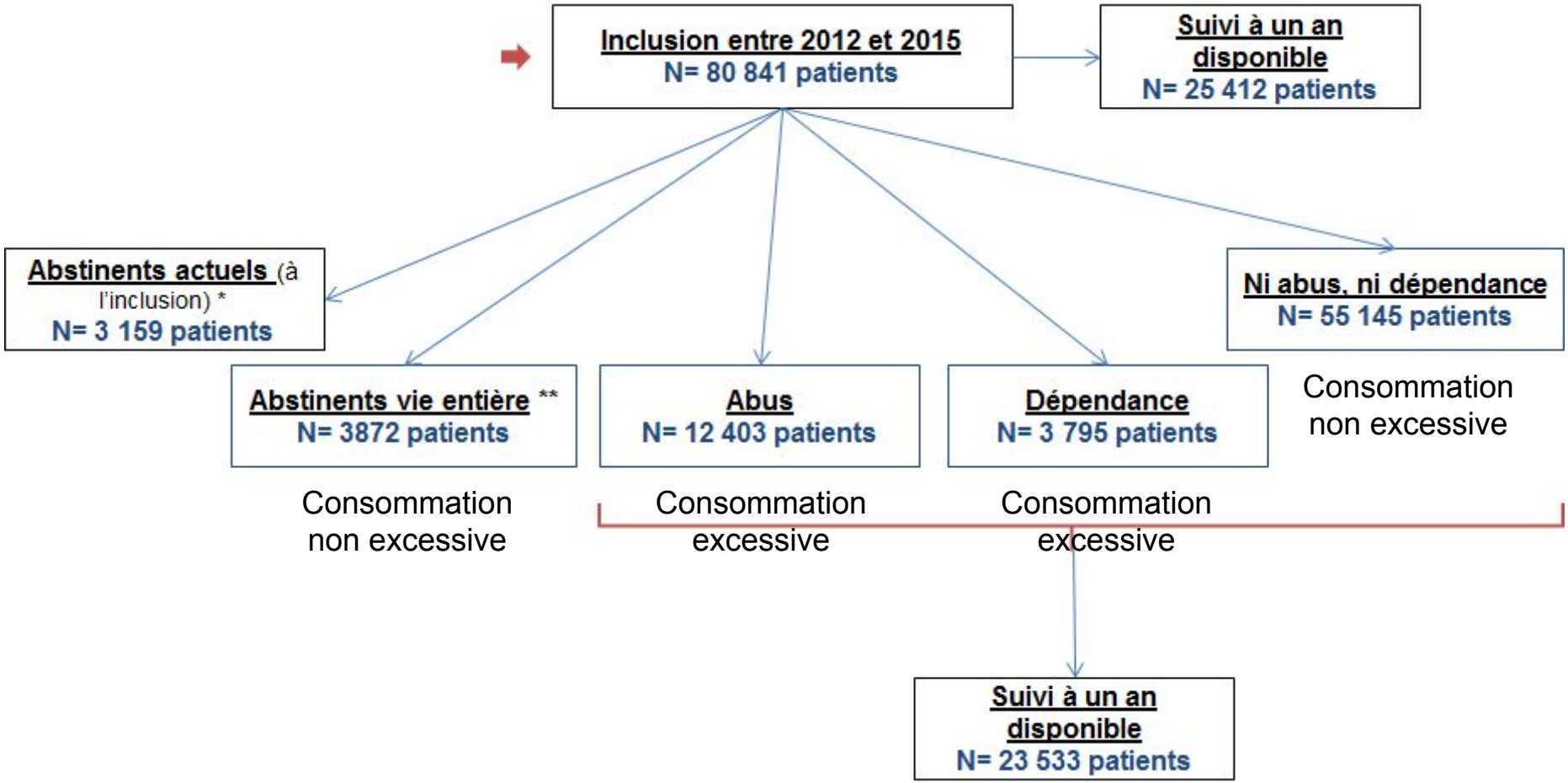
Contexte : le SNIIRAM

- ➔ SNIIRAM : Données de l'assurance-maladie de l'ensemble de la population française
- ➔ Données issues des prestations remboursées
  - ◆ Consultations médicales
  - ◆ Remboursements de médicaments
  - ◆ Hospitalisations
  - ◆ Affections Longue Durée
  - ◆ Examens biologiques
  - ◆ Autres prestations remboursées par la sécurité sociale
- ➔ Pour le stage:
  - ◆ Diagnostics d'hospitalisations
  - ◆ Consommations de médicaments
  - ◆ Accidents graves du travail

Consommation d'alcool

- ➔ Score AUDIT calculé à partir du questionnaire
- ➔ 10 items
- ➔ 3 classes selon la consommation
  - ◆ Ni abus ni dépendance (Hommes : score < 8, femmes : < 7)
  - ◆ Abus (H : [8,12] , F : [7,11])
  - ◆ Dépendance (H : >12, F : >11)


Inclusion entre 2012 et 2015
N= 80 841 patients
Suivi à un an disponible
N= 25 412 patients
Abstinents actuels (à l'inclusion) *
N= 3 159 patients
Ni abus, ni dépendance
N= 55 145 patients
Abstinents vie entière **
N= 3872 patients
Abus
N= 12 403 patients
Dépendance
N= 3 795 patients
Suivi à un an disponible
N= 23 533 patients


Inclusion entre 2012 et 2015
N= 80 841 patients
Suivi à un an disponible
N= 25 412 patients
Abstinents actuels (à l'inclusion) *
N= 3 159 patients
Ni abus, ni dépendance
N= 55 145 patients
Abstinents vie entière **
N= 3872 patients
Abus
N= 12 403 patients
Dépendance
N= 3 795 patients
Consommation non excessive
Consommation non excessive
Consommation excessive
Consommation excessive
Suivi à un an disponible
N= 23 533 patients

Contexte

- ➔ But : évaluer l'effet de la consommation d'alcool
- ➔ Problème statistique : équilibrer les données entre 2 groupes pour limiter les biais de confusion
- ➔ Idéal : **randomisation**
  - ◆ Impossible ici -> Stratification ? Appariement ? Ajustement ?

But du score de propension à haute dimension

Deux problèmes :

- ➔ Équilibrer les données
- ➔ Très grand nombre de variables

100 variables binaires : 2^{100} combinaisons possibles…

- ➔ Difficile de stratifier ou de valider les modèles

Le score de propension

Le score est la probabilité de recevoir le traitement, sachant les covariables.

$$e(x_i) = P(W_i = 1 | X_i)$$

Principal avantage : **réduire n covariables en 1 seule dimension**

La distribution des covariables est similaire pour deux individus ayant le même score

Le score de propension à haute dimension

- ➔ Un trop grand nombre de variables : nécessité d'effectuer une sélection
- ➔ Pour cette étude : 7133 diagnostics et 1194 médicaments = 8327 variables binaires

Sélection des variables

- ➔ Les variables issues de CONSTANCES restent quoi qu'il arrive dans le modèle
- ➔ On sélectionne parmi les tables DIAGNOSTICS et MEDICAMENTS
- ➔ On cherche à garder les variables liées à la fois à l'exposition et à la réponse. On utilise la formule de Bross pour calculer le biais multiplicatif de chaque variable

$$Bias_M = \frac{P_{C_1}(RR_{CD}-1)+1}{P_{C_0}(RR_{CD}-1)+1}$$

P_{C_1} : Fréquence d'apparition de la variable parmi le groupe de consommation excessive d'alcool

P_{C_0} : Fréquence d'apparition de la variable parmi le groupe de consommation non excessive d'alcool

RR_{CD} : Association entre la covariable et la variable réponse, mesurée en terme de risque relatif

Sélection des variables

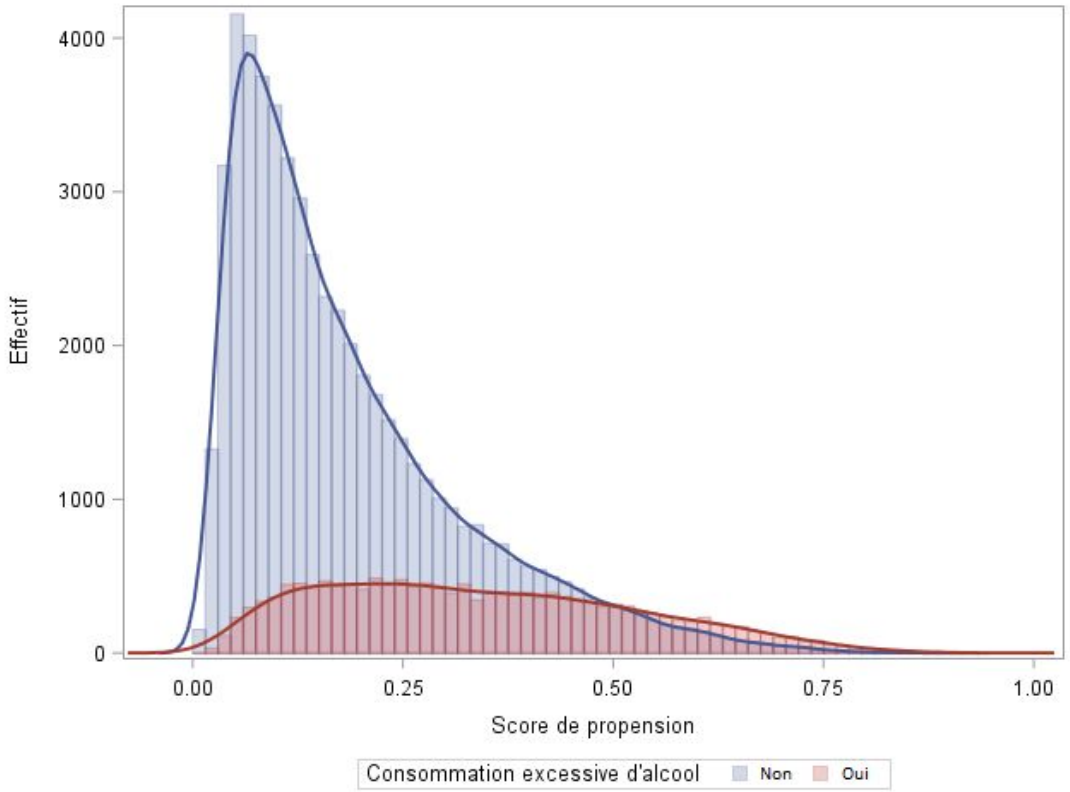
- ➔ Formule de Bross calculée sur les variables présentes pour au moins 1% de la population
- ➔ On garde les 75% meilleurs scores

Calcul du score

- ➔ Régression logistique sur l'ensemble des variables mesurées
- ➔ Le score est compris entre 0 et 1.

Calcul du score

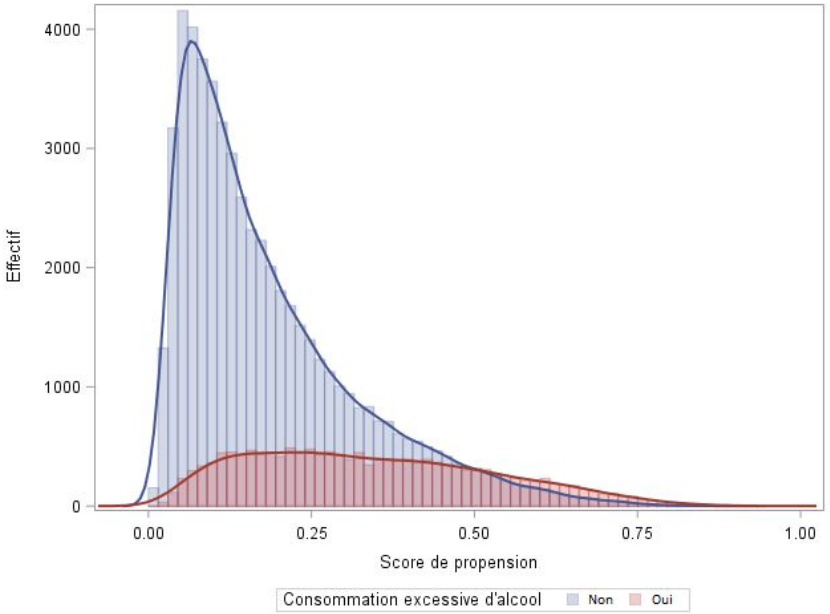
➔ Zone de support commun suffisamment large

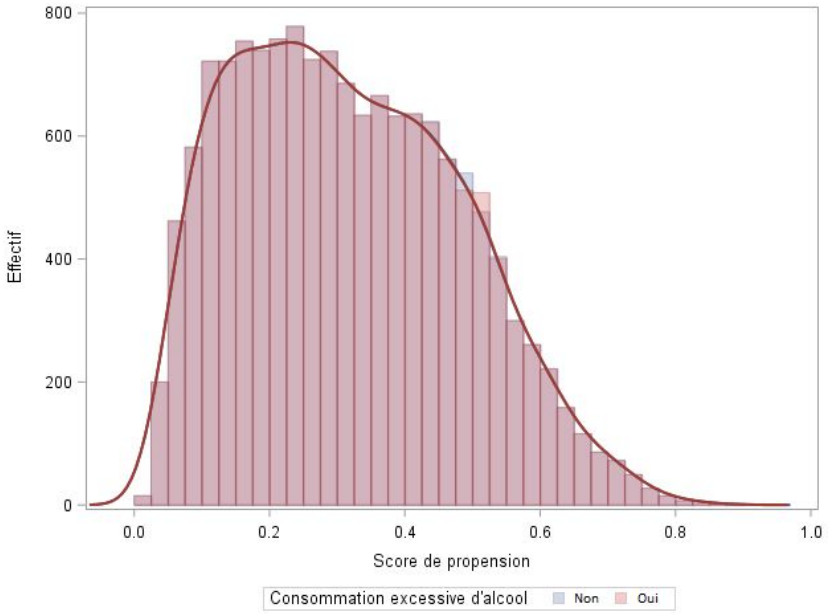


Appariement

- ➔ Comment utiliser le score ?
  - ◆ Pondération
  - ◆ Ajustement
  - ◆ **Appariement**
- ➔ Méthode des plus proches voisins
- ➔ Appariement par paires selon le score.
- ➔ Écart maximal (caliper) : $0.25 * \sigma(e(X))$ soit 0.04

Appariement





Appariement

- ➔ Equilibrage obtenu sur la distribution du score
- ➔ Equilibrage obtenu sur l'ensemble des variables issues de la cohorte CONSTANCES

Les populations sont maintenant comparables !

Résultats

- ➔ Données équilibrées : **l'ajustement n'est plus nécessaire**
  - ◆ **On peut comparer directement les deux populations**

- ➔ En pratique… Equilibre imparfait
  - ◆ Un nouvel ajustement peut être utile

Résultats

Plusieurs modèles :

1. Sans ajustement
2. Ajustement sur le score
3. Ajustement sur les variables insuffisamment équilibrées

A comparer au modèle 0 sur la population non appariée

Résultats

Accident entre 2012 et 2016	Consommation excessive	Consommation non excessive	Total
Oui	225 (1.4%)	860 (1.5%)	1085 (1.4%)
Non	58157 (98.5%)	15973 (98.6%)	74130 (98.6%)

Modèle	Odds-ratio Consommation excessive VS consommation non excessive	Intervalle de confiance	p-val Khi-2
1	0.946	0.778-1.152	0.5821
2	0.946	0.778-1.152	0.5818
3	0.947	0.778-1.152	0.5847

Résultats

Dépressivité	Consommation excessive	Consommation non excessive	Total
Oui	4212 (27.2%)	11675 (21.1%)	15887 (22.4%)
Non	11278 (72.8%)	43692 (78.9%)	54970 (77.6%)
Total	55367	15490	70857
Valeurs manquantes : 4358			

Modèle	Odds-ratio Consommation excessive VS consommation non excessive	Intervalle de confiance	p-val Khi-2
1	1.230	1.166-1.298	<.0001
2	1.231	1.167-1.299	<.0001

Conclusion

- ➔ Pas de nouveau résultat obtenu mais…
- ➔ Bonne méthode d'**équilibrage des données** (ne remplace pas la randomisation !)
- ➔ Confirmation de l'effet de l'alcoolisme sur la dépressivité
- ➔ Accidents du travail : résultat inattendu, travail à poursuivre

- ➔ Utilisation possible sur d'autres variables réponses, avec d'autres modèles