Emulation d'essais contrôlés randomisés à partir de données observationnelles

Raphaël Porcher

CRESS, Université Paris Descartes, Inserm UMR-S 1153

GDR Stat Santé
Nantes, 28 septembre 2018

Plan

Effet « causal »

- Effet d'un « traitement » A sur un critère de réponse Y
- « Traitement » : toute intervention ou exposition
  - Que l'on peut mettre en oeuvre ou éviter (médicament, type de greffe, tabac, pollution . . .)
  - Pas un état non modifiable (sexe, maladie, . . .)
- Effet en comparaison d'une intervention contrôle (ou abstention)
- C'est le but des essais contrôlés randomisés (ECR) !
- Mais parfois on peut aussi utiliser des données observationnelles

Effet du traitement et ECR

- Dans l'idéal, on aurait $Y(1)$ et $Y(0)$ et effet causal = $Y(1) - Y(0)$
- On teste $\mathrm{E}[Y(1) - Y(0)] = 0$ (pas d'effet causal en moyenne)
- Logique de l'essai croisé (*cross-over*)
  - Souvent difficile à mettre en place
  - Et on n'observe pas vraiment $Y(1)$ et $Y(0)$
- On suppose toutefois que $Y(1)$ et $Y(0)$ existent (modèle contre-factuel de Rubin)
- ECR : allocation aléatoire de A
  - D'un point de vue math, $(Y(1), Y(0)) \perp A$
  - $\mathrm{E}[Y(a)|A = 1] = \mathrm{E}[Y(a)|A = 0] = \mathrm{E}[Y(a)]$, $a \in \{0, 1\}$
  - $\mathrm{E}[Y(1)|A = 1] - \mathrm{E}[Y(0)|A = 0] = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] = \mathrm{E}[Y(1) - Y(0)]$

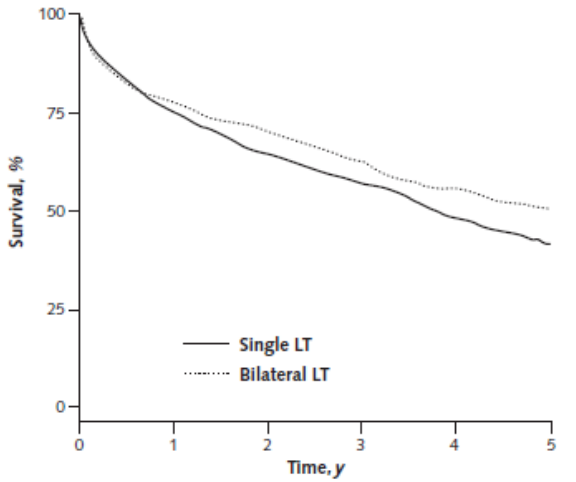
ECR pas toujours faisable

- La randomisation est parfois non éthique
  - Par exemple tabac, toxicomanie, violence, . . .
  - Ou considérée comme un frein à la participation des patients
- Parfois pas faisable
  - Croyance des soignants (cas d'une technique chirurgicale ou d'une prothèse déjà utilisée, . . .)
  - Parce que l'on veut étudier ce qui se passe en vie réelle ($\neq$ expérience)
  - Parce que l'on ne peut pas comparer toutes les alternatives thérapeutiques dans tous les sous-groupes avec des ECR

$\rightarrow$ Études observationnelles

Effet du traitement dans une étude observationnelle

- On ne contrôle plus les sources de biais $\rightarrow$ biais d'indication
- Facteurs de confusion
  - $Y(a)$ dépend de X, et A aussi
  - Donc $(Y(1), Y(0))$ n'est plus indépendant de A
  - On peut estimer $\mathrm{E}[Y(a)|A = a)]$ mais plus $\mathrm{E}[Y(a)]$
- Exemple : transplantation mono- *vs* bi-pulmonaire pour des fibroses pulmonaires idiopathiques



At risk, *n*						
Single LT	2146	1378	1003	765	551	404
Bilateral LT	1181	648	435	284	199	129
Total	3327	2026	1438	1049	750	533

- Une transplantation bipulmonaire entraîne-t-elle une meilleur survie ?
- Ou n'a-t-on réalisé cette intervention plus lourde que chez des patients en meilleur état général ?

Solutions

- Méthodes d'analyse "classiques"
  - Stratification ou appariement (difficile si beaucoup de variables)
  - Modèle de régression (hypothèses, problèmes de dimensionnalité, problèmes d'extrapolation)
- Méthodes dites d'inférence causale
  - Scores de propension
  - Variables instrumentales
  - D'autres approches plus rares ou plus complexes

Score de propension

- Probabilité de recevoir A sachant les caractéristiques X

$$e(X) = \Pr(A|X)$$

- Propriétés des SP : sous certaines hypothèses
  - Le SP est un score d'équilibrage
  - Conditionner sur le SP conduit à un estimateur sans biais de l'effet du traitement
  - Utiliser une estimation de $e(X)$ conserve ces propriétés

- Hypothèses
  - *Stable Unit Value Assumption* (SUTVA) :
    - L'allocation du traitement d'un patient ne modifie pas les réponses potentielles des autres : $(Y_i(1), Y_i(0), A_i)$ sont indépendantes
    - Pas de versions « cachées » du traitement : si $A = a$, on observe $Y(a)$
  - Tous les facteurs de confusion sont mesurés
  - Positivité : $0 < \Pr(A = 1|X) < 1, \forall X$
  - Les deux précédentes = ignorabilité (SITA) : $(Y(1), Y(0)) \perp A|X$

Comment la théorie « fonctionne » ?

1. SITA : $\mathrm{E}_X\left[\mathrm{E}(Y|A=a,X)\right] = \mathrm{E}_X\left\{\mathrm{E}[Y(a)|A=a,X]\right\} = \mathrm{E}_X\left\{\mathrm{E}[Y(a)|X]\right\} = \mathrm{E}[Y(a)]$
2. Score d'équilibrage : $(Y(1),Y(0))\perp A|X \rightarrow (Y(1),Y(0))\perp A|e(X)$

- Donc, conditionner sur $e(X)$ suffit et permet d'estimer $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$
- Cela reste vrai en utilisant $\hat{e}(X)$ plutôt que $e(X)$ (que l'on ne connaît pas)

Étapes-clé d'une analyse par SP

1. Définir l'intervention et la population-cible
2. Identifier les données appropriées (étude ad-hoc ou BDD existante)
3. Sélectionner les variables nécessaires (facteurs de confusion)
4. Estimer le SP (modèle logistique, CART, forêts aléatoires . . .)
5. « Conditionner » sur le score de propension
6. Évaluer le déséquilibre résiduel entre les groupes (succès du SP)
7. Analyser le critère de jugement

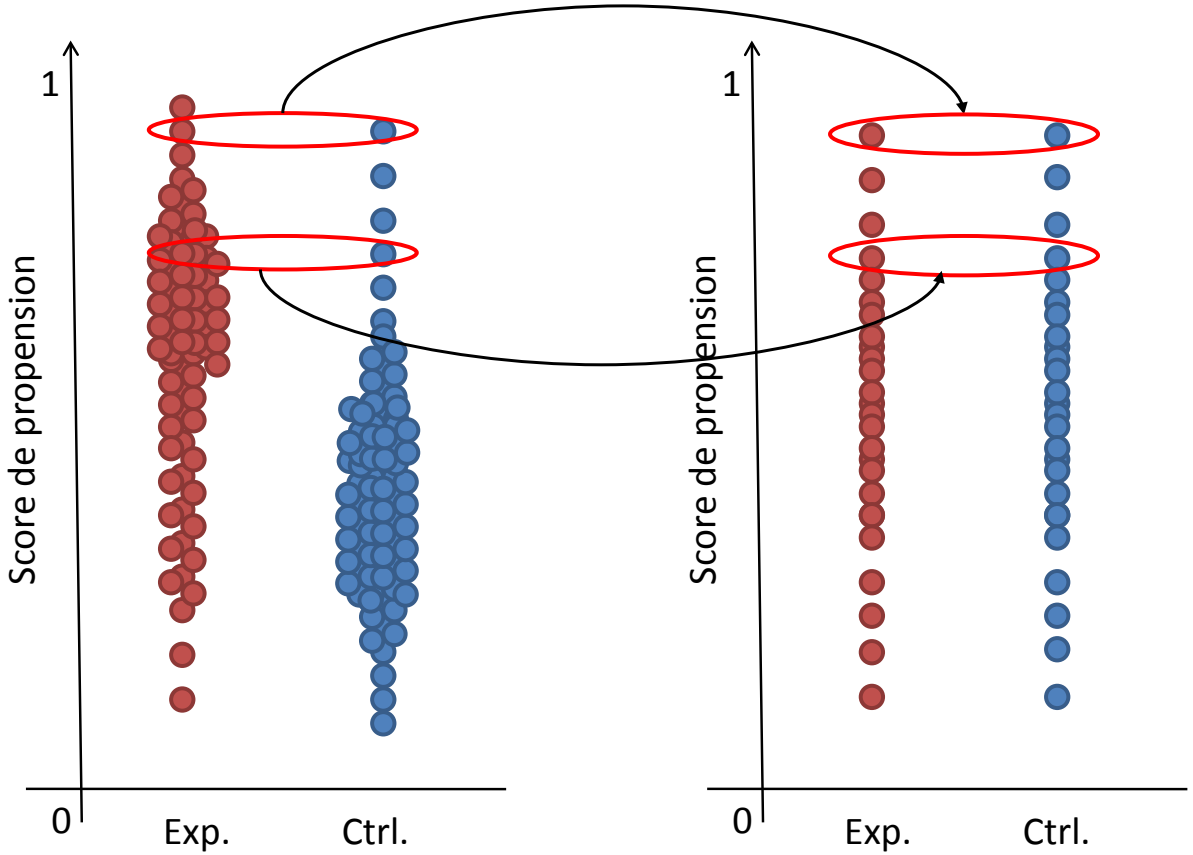
« Conditionner » sur le score de propension

- En pratique, « conditionner »peut conduire à plusieurs analyses
- Ajustement (modèle de régression) peu efficace
- Stratification (sur des classes de SP)
- Appariement
- Pondération

Appariement

- Apparier k témoins à 1 traité (le plus souvent $k = 1$)
- Apparier au plus proche voisin rend toujours un résultat, mais pas peut conduire à des bias (le plus proche peut être éloigné)
- Le plus souvent : appariement dans un intervalle (*caliper*) de SP
  - Usuellement 0.2 ET du SP
  - Retire $\geq$ 98% du bais (résultat empirique)
- Méthode critiquée récemment (Gary King, 2016), mais qui reste la plus populaire

Appariement (II)



Avant appariement
(données initales)

Après appariement

Exemple : transplantation mono- *vs* bi-pulmonaire [1]

- Patients avec une fibrose pulmonaire idiopathique
- Intervention = BLT *vs* SLT
- Critère de jugement = survie
- Registre UNOS, 3327 patients
- Appariement 1 : 1 sans remise, caliper de 0.25 ET

1. Thabut et coll. 2009

Données initiales

Table 1. **Main Baseline Patient Characteristics, by Type of Lung Transplantation**

Characteristic	Nonmissing Data, *n* (%)	Single-Lung Transplantation (*n* = 2146)	Bilateral Lung Transplantation (*n* = 1181)	Standardized Difference, %*
Recipient				
Mean age (SD), *y*	3327 (100)	57.1 (9.0)	54.0 (10.0)	32.1
Age distribution, *n* (%)	3327 (100)			
≤50 y		424 (19.8)	362 (30.7)	25.3
51–55 y		324 (15.1)	188 (15.9)	2.3
56–60 y		552 (25.7)	279 (23.6)	4.9
>60 y		846 (39.4)	352 (29.8)	20.3
Women, *n* (%)	3327 (100)	705 (32.9)	358 (30.3)	5.5
Functional status, *n* (%)†	2852 (85.7)			
Class I		464 (26.0)	213 (19.9)	14.6
Class II		928 (52.1)	483 (45.1)	13.9
Class III		390 (21.9)	374 (35.0)	29.3
Diabetes, *n* (%)	3044 (91.5)	279 (14.7)	186 (16.2)	4.1
Oxygen required at rest, *n* (%)	2498 (75.1)	1318 (76.1)	642 (83.8)	19.4
Mean FVC (SD), % *predicted*	3082 (92.6)	49.0 (16.0)	47.4 (17.9)	9.3
Mean pulmonary capillary wedge pressure (SD), *mm Hg*	2842 (85.4)	8.8 (5.9)	10.1 (6.1)	22.8
Mean pulmonary artery pressure (SD), *mm Hg*	2474 (74.4)	23.4 (8.8)	28.4 (11.5)	49.2
Mean body mass index (SD), kg/m^2	3193 (96.0)	27.2 (4.5)	26.8 (4.3)	11.0
Donor				
Mean age (SD), *y*	3327 (100)	32.2 (13.6)	33.0 (14.9)	5.3
Female, *n* (%)	3327 (100)	775 (36.1)	532 (45.0)	18.3
Mean body mass index (SD), kg/m^2	3146 (94.6)	24.8 (5.1)	25.0 (5.1)	3.2
Diabetes, *n* (%)	3060 (91.9)	76 (4.0)	46 (4.0)	0
Cause of death, *n* (%)	3149 (94.6)			
Anoxia		136 (6.8)	98 (8.5)	6.3
Stroke		740 (37.1)	445 (38.6)	3.0
Head trauma		1105 (55.4)	599 (51.9)	7.0
CNS tumor		14 (0.7)	12 (1.0)	3.6
Donor-to-recipient				
Cytomegalovirus status mismatches, *n* (%)	2361 (71.0)	610 (44.3)	434 (44.2)	0.2
Sex mismatches, *n* (%)	3327 (100)	616 (28.7)	418 (35.4)	14.4
Blood group mismatches, *n* (%)	3327 (100)	221 (10.3)	101 (8.6)	6.0
HLA mismatches, *n* (%)	2735 (82.2)	4.6 (1.1)	4.7 (1.1)	7.6

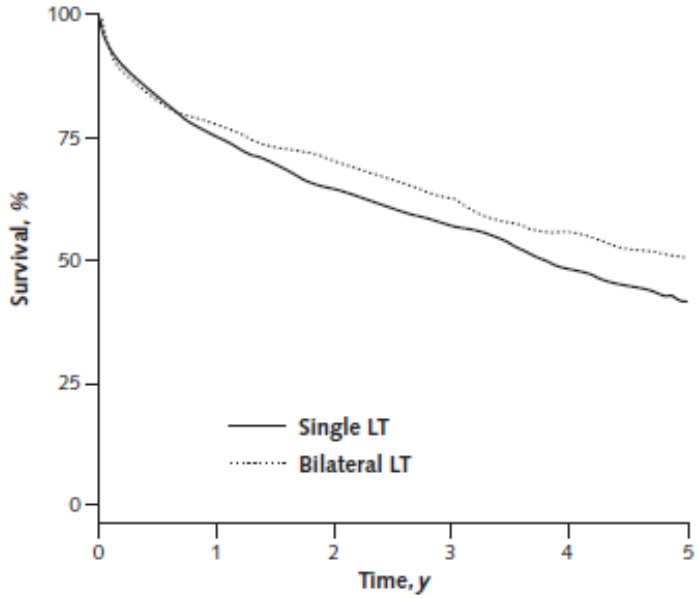
CNS = central nervous system.
* Mean difference divided by the pooled SD, expressed as a percentage.
† Ranges from class I to III, indicating that the patient performs activities of daily living with no, some, or total assistance, respectively.

Données appariées

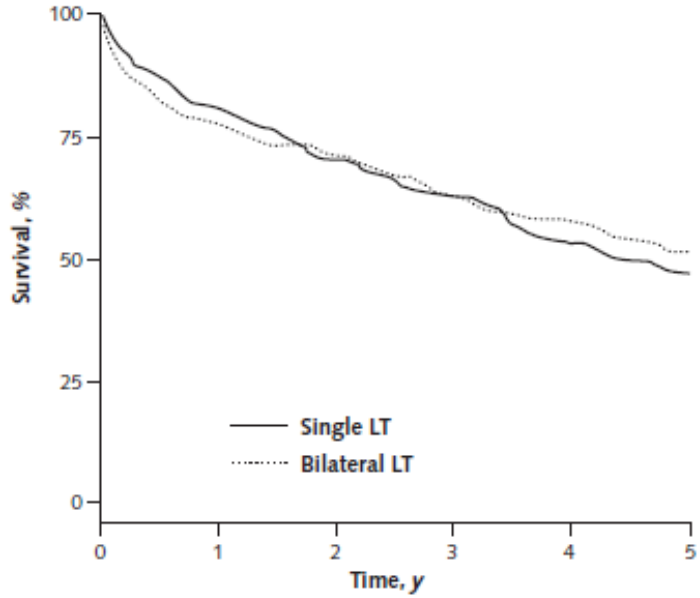
Table 2. **Main Baseline Characteristics of Patients Matched by Propensity Score, by Type of Lung Transplantation**

Characteristic	Single-Lung Transplantation (n = 795)	Bilateral Lung Transplantation (n = 795)	Standardized Difference, %*
Recipient			
Mean age (SD), *y*	56.0 (8.4)	55.9 (8.4)	0.7
Age distribution, *n (%)*			
≤50 y	172 (21.6)	180 (22.6)	2.4
51–55 y	134 (16.9)	126 (15.8)	2.7
56–60 y	218 (27.4)	214 (26.9)	1.1
>60 y	271 (34.1)	275 (34.6)	1.1
Women, *n (%)*	244 (30.7)	229 (28.8)	4.2
Functional status, *n (%)*†			
Class I	179 (22.5)	173 (21.8)	1.8
Class II	369 (46.4)	376 (47.3)	1.8
Class III	247 (31.1)	246 (30.9)	0.3
Diabetes, *n (%)*	143 (18.0)	125 (15.7)	6.0
Oxygen required at rest, *n (%)*	674 (84.8)	672 (84.5)	0.7
Mean FVC (SD), *% predicted*	48.9 (16.6)	48.5 (17.4)	2.4
Mean PCWP (SD), *mm Hg*	9.7 (6.0)	9.5 (5.6)	4.5
Mean pulmonary artery pressure (SD), *mm Hg*	24.8 (8.6)	24.7 (8.7)	0.3
Body mass index (SD), kg/m^2	27.2 (4.4)	26.9 (4.2)	5.8
Donor			
Mean age (SD), *y*	32.9 (13.8)	33.3 (15.0)	2.5
Female, *n (%)*	334 (42.0)	329 (41.4)	1.3
Mean body mass index (SD), kg/m^2	25.0 (5.2)	25.0 (5.1)	0.2
Diabetes, *n (%)*	31 (3.9)	36 (4.5)	2.0
Cause of death, *n (%)*			
Anoxia	64 (8.1)	68 (8.6)	1.8
Stroke	297 (37.4)	305 (38.4)	2.1
Head trauma	425 (53.5)	412 (51.8)	3.3
CNS tumor	9 (1.1)	10 (1.3)	1.2
Donor-to-recipient			
Cytomegalovirus status mismatches, *n (%)*	363 (45.7)	354 (44.5)	2.3
Sex mismatches, *n (%)*	248 (31.2)	248 (31.2)	0
Blood group mismatches, *n (%)*	75 (9.4)	74 (9.3)	0.4
HLA mismatches, *n (%)*	4.6 (1.0)	4.6 (1.1)	3.4

Analyse du critère de jugement



At risk, *n*						
Single LT	2146	1378	1003	765	551	404
Bilateral LT	1181	648	435	284	199	129
Total	3327	2026	1438	1049	750	533

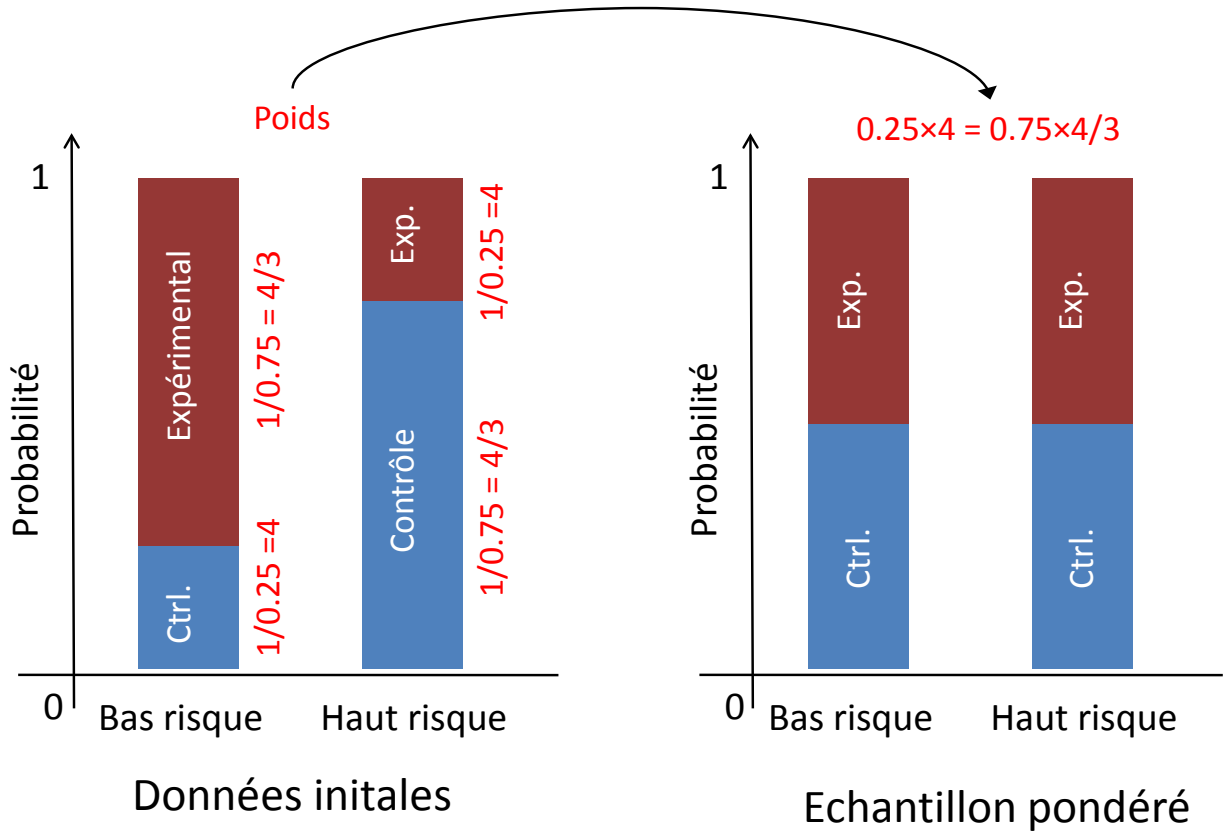


At risk, *n*						
Single LT	795	499	333	235	153	112
Bilateral LT	795	456	325	225	164	105
Total	1590	955	658	460	317	217

Inverse probability of treatment weighting (IPTW)

- Idée : reconstruire une population de structure similaire dans les feu groupes
- Patients traités pondérés par $1/\hat{e}(X_i)$
- Patients témoins pondérés par $1/[1-\hat{e}(X_i)]$
- Sur-pondère les patients qui avaient une probe faible de recevoir leur traitement
  - Compense le plus grand nombre de patients avec le même $e(X)$ dans l'autre bras
- D'autres types de pondération sont possibles pour estimer d'autres effets (pas le sujet aujourd'hui)

IPTW


Poids
0.25×4 = 0.75×4/3
1
Probabilité
0
Expérimental
Ctrl.
Exp.
Contrôle
1/0.75 = 4/3
1/0.25 =4
1/0.25 =4
1/0.75 = 4/3
Bas risque
Haut risque
Données initales
1
Probabilité
0
Exp.
Ctrl.
Exp.
Ctrl.
Bas risque
Haut risque
Echantillon pondéré

Quel est le problème ?

- Exemple de la TP : à un instant donné, une intervention (TP), et choix entre SLT et BLT
- Situation rare avec des donnés de vie réelle ou d'une étude de cohorte
- Situation plus classique : patients suivis et à un moment du suivi, l'intervention est appliquée à un(e) patient(e)
- Exemples :
  - Cohorte : Traitement préventif par IEC d'enfants atteints de myopathie de Duchenne (travail en cours)
  - BDD médico-administrative : Dépistage du cancer colorectal par coloscopie (Garcia-Albeniz et al.)
- On identifie bien les patients traités, mais qui sont les contrôles ? Quand ?
- Si on prend comme contrôles ceux qui ne sont jamais traités $\rightarrow$ biais d'immortalité

Possibles solutions

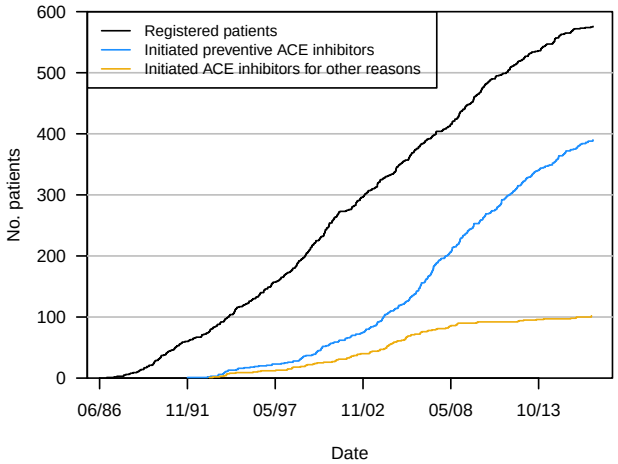
- Considérer le traitement comme une exposition qui varie dans le temps et utiliser une approche analytique adapté
  - *g*-formula
  - MSM (IPTW)
  - SNM
  - IPCW (pour censure non-indépendante / informative)
- Essayer d'émuler un essai clinique tel qu'on aurait pu le faire
  - Et tout de même utiliser des méthodes citées ci-dessus (*g*-methods)

g-formula, IPTW, MSM

- On a maintenant A(t) et X(t) avec $t = 0, \ldots, K$ pour simplifier (extension possible au temps continu)
- Histoire : $\overline{A}(t) = \{A(0), A(1), \ldots, A(t)\}$, idem $\overline{X}(t)$
- Prenons un des 2^{K+1} régimes $\bar{a} = \{a(0), \ldots, a(K)\}$ et on cherche $\mathrm{E}(Y_{\bar{a}})$
- *g*-formula : $\hat{\mathrm{E}}(Y_{\bar{a}}) = \sum_{\bar{x}} \mathrm{E}(Y|\overline{A} = \bar{a}, \overline{X} = \bar{x}) \prod_{t=0}^{K} f[x(t)|\overline{A}(t-1) = \bar{a}(t-1), \overline{X}(t-1) = \bar{x}(t-1)]$
- IPTW : estimer $\mathrm{E}(Y_{\bar{a}})$ come la moyenne pondérée de Y parmi ceux avec $\overline{A} = \bar{a}$, poids $w^* = \prod_{t=0}^{K} \frac{f[a(t)|\bar{a}(t-1)]}{f[a(t)|\bar{a}(t-1),\bar{x}(t)]}$
- 2^{K+1} très grand $\rightarrow$ utiliser de spécifier un modèle non saturé pour $\mathrm{E}(Y_{\bar{a}})$ = MSM

Retour à l'exemple DMD

- Intervention : IEC à titre préventif (i.e. avec FE préservées $\neq$ curatif)
- Données = cohorte d'enfants inclus entre 1986 et 2017



- On pourrait supposer un MSM et estimer ses paramètres avec ces données
- Mais si on avait fait un essai, ça n'aurait jamais été avant 1992, donc cette partie des données ne devrait pas servir, la question des IEC à titre préventif ne se pose pas pour les patients qui reçoivent des IEC à titre curatif (en jaune), . . .

$\rightarrow$ Essai émulé

Planifier un essai . . . et l'émuler

Protocole de l'essai cible	Émulation de l'essai cible
Critères d'éligibilité	Critères d'éligibilité
Interventions	Interventions
Randomisation	« Mimer » la randomisation
Début/fin de suivi	Début/fin de suivi
Critères de jugement	Critères de jugement
Contraste causal	Contraste causal
Plan d'analyse	Plan d'analyse

- Permet de définir précisément la population d'étude
- De limiter les biais de confusion
- D'éviter les biais d'immortalité
- De bien spécifier les contrastes que l'on veut étudier

Population d'étude

- Exemple DM :
  - Myopathie de Duchenne
  - Âge entre 8 et 13 ans, entre 01/1992 et 12/2016
  - Non traité par IEC
  - Avec une FE $\geq$ 55%
- Exemple coloscopie
  - Âge 70–74 ans en 2004-2012
  - Pas d'ATCD de MICI, adénome, colotomie, dépistage dans les 5 ans précédents
  - Pas de symptômes gastro-intestinaux les 6 derniers mois
  - Participant à Medicare depuis au moins 5 ans

Temps d'origine (*time zero*)

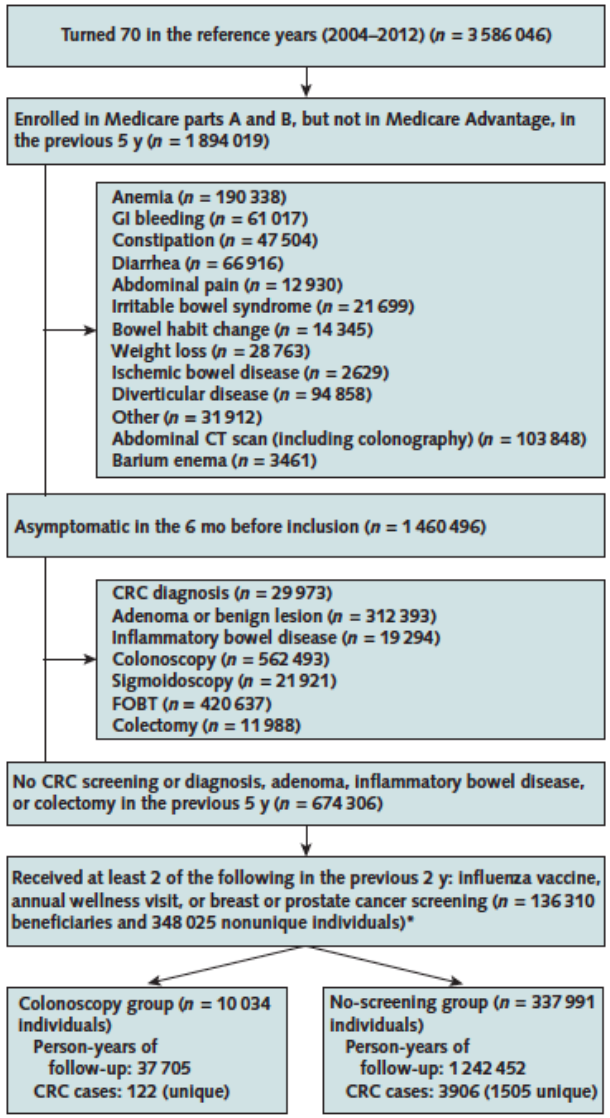
- Dans un ECR, les patients sont inclus (et randomisés) à une date donnée
  - Critères d'inclusion vérifiés
  - Traitement alloué par tirage au sort
  - Début du suivi pour le critère de jugement
- Dans un essai émulé, il faut reproduire cela
- Pour éviter des biais d'immortalité
- Temps zéro
  - On pourrait considérer la date de traitement pour les traités (Cf infra)
  - Et la date où les patients non-traités vérifient les critères d'inclusion
  - Mais un même patient peut vérifier les critères d'inclusion plusieurs fois

Choix d'un temps d'origine

- Un seul temps zéro
  - Date de traitement pour les traités (si éligibles)
  - Première date où les critères sont remplis pour un contrôle
  - Ou un de ces temps choisis au hasard
- Prendre tous les temps éligibles
  - Émuler une séquence d'essais, un par unité de temps (1 semaine pour coloscopie, 3 mois pour DMD)
  - Pour chaque essai (émulé), inclure tous les patients éligibles (ou un sous-groupe aléatoire des contrôles, 5% pour coloscopie, par exemple)
  - Combiner la séquence d'essais émulés
  - Tenir compte des mêmes individus inclus plusieurs fois avec une variance robuste ou bootstrap

Exemple coloscopie

Figure 1. Flow into colonoscopy screening groups of Medicare beneficiaries aged 70 years, 2004–2012.



CRC = colorectal cancer; CT = computed tomography; FOBT = fecal occult blood testing; GI = gastrointestinal.

Tenir compte des biais de sélection / confusion

- De nombreuses approches sont possibles
- Coloscopie : proche de *g*-formula
  - Estimer un modèle avec bras, covariables à l'inclusion et le temps de suivi (linéaire, quadratique + décroissance exponentielle) et interactions temps $\times$ bras
  - Prédire l'incidence de cancer colorectal pour chacun des sujets sous chacune des intervention
  - Moyenner ces prédiction sur l'ensemble des patients
- Myopathie de Duchenne : IPTW

Contrastes d'intérêt

- Coloscopie : comparer l'incidence pour les patients ayant eu une coloscopie à ceux n'en ayant pas eu (dans l'essai émulé). *Treatment initiators*
- Myopathie de Duchenne : on peut aussi faire cela $\rightarrow$ mimer une analyse en ITT
- Mais 91% des contrôles reçoivent des IEC, 72% à titre préventif . . .
- Choix de comparer les patients traités à ceux non traités : censure des contrôle quand ils reçoivent des IEC à titre préventif, et IPCW [2]
- D'autres contrastes sont possibles, p.ex. per protocol (censurer aussi à l'arrêt des IEC pour les traités), . . .

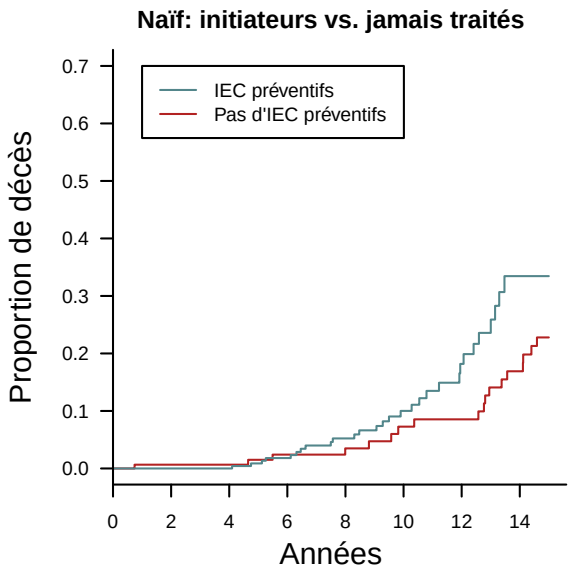
2. idem stratification séquentielle, Schaubel et al. *JASA* 2009

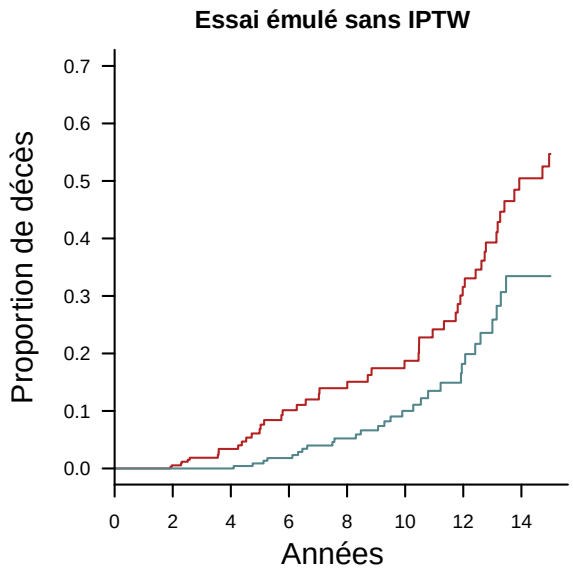
Autres points à considérer

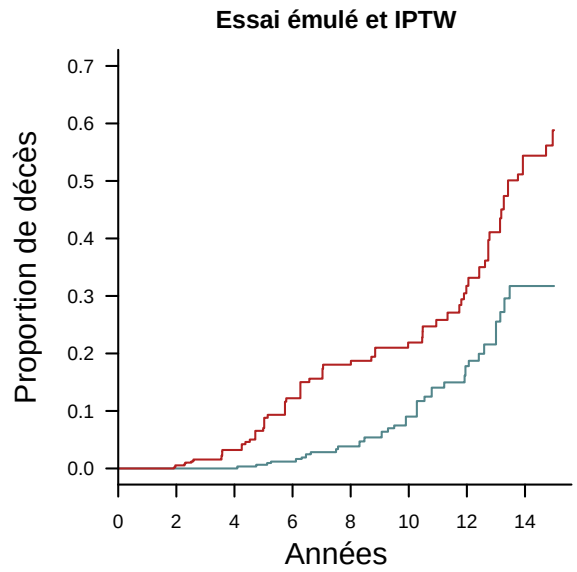
- Prendre en compte une « période de grâce »
  - Par exemple laisser 3 mois pour initier le traitement après inclusion
  - Si un patient est éligible en début de période et meurt dans les deux mois sans être traité, alors il(elle) aurait pu être alloué au traitement et ne pas l'avoir pris
  - Trier au sort le bras, ou dupliquer l'individu
- Myopathie de Duchenne : le critères d'éligibilité nécessitent de connaître la FE tous les trois mois
  - Les données ne contiennent pas de mesures si rapprochées
  - On utilise le BLUP obtenu à partir d'un modèle mixte flexible [3]

3. Cf. Taylor et al. *Stat Med* 2014

Résultats préliminaires DMD







Conclusions

- Essai émulé : pas tant un modèle statistique qu'un cadre pour aborder une question causale
- Différentes méthodes de la statistique dite causale peuvent ensuite être utilisées
- Ne remplacera jamais un essai randomisé
- Mais il y a de nombreux cas où il n'y a pas d'ECR ou on ne peut pas en faire
- Améliore certains défauts des études observationnelles
  - Limite les biais de sélection et d'immortalité
  - Les questions liées aux facteurs de confusion non mesurés restent . . .
- Ne pas oublier que toute analyse autre qu'ITT d'un ECR utilise aussi des données observationnelles et nécessite des méthodes d'inférence causale (p. ex. IPCW, CACE, . . .)

Merci de votre attention !

. . . et à Karim Wahbi (cardiologie, Hôpital Cochin), Gabriel Thabut (pneumologie, Hôpital Bichat), Justine Jacot (CRESS)

Quelques références


Danaei G, Rodríguez LA, Cantero OF, Logan R, Hernán MA. Observational data for comparative effectiveness research : an emulation of randomised trials of statins and primary prevention of coronary heart disease. *Stat Methods Med Res* 2013 ; 22(1) :70–96


García-Albéniz X, Hsu J, Bretthauer M, Hernán MA. Effectiveness of screening colonoscopy to prevent colorectal cancer among Medicare beneficiaries aged 70 to 79 years. *Ann Intern Med* 2017 ; 166(1) :18–26.


Hernán M, Brumback B, Robins J. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J Am Stat Assoc* 2001 ; 96 :440–448.


Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol.* 2016 ; 183(8) :758–764.


Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects : a comparative study. *Stat Med* 2004 ; 23(19) : 2937–2960.


Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000 ; 11(5) :550–560.


Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983 ; 70 :41–55.


Schaubel DE, Wolfe RA, Sima CS, et al. Estimating the effect of a time-dependent treatment by levels of an internal time-dependent covariate : application to the contrast between liver wait-list and posttransplant mortality. *J Am Stat Assoc* 2009 ; 104(485) :49– 59.


Taylor JM, Shen J, Kennedy EH, et al. Comparison of methods for estimating the effect of salvage therapy in prostate cancer when treatment is given by indication. *Stat Med* 2014 ; 33(2) :257–274.


Thabut G, Christie JD, Ravaud P, Castier Y, Dauriat G, Jebrak G, Fournier M, Leseche G, Porcher R, Mal H. Survival after bilateral versus single-lung transplantation for idiopathic pulmonary fibrosis. *Ann Intern Med* 2009 ; 151 :767–774.