Estimation de l'incidence départementale des cancers en France : comment et jusqu'où les big-data peuvent-elles contribuer ?

Z Uhry[1,2,3], E Chatignoux[2], M Colonna[3,4], N Bossard[1,3], L Remontet[1,3]

[1] Service de Biostatistique, Hospices Civils de Lyon
[2] Santé Publique France
[3] Université Lyon1 / CNRS, UMR 5588- Laboratoire de Biométrie & Biologie Evolutive
[4] Registre du cancer de l'Isère
[5] Réseau des registres français du cancer Francim

Forum de la recherche en cancérologie, 5 avril 2017
CLARA - Cancéropôle Auvergne-Rhône-Alpes

Contexte

- Les registres départementaux de cancers enregistrent de façon exhaustive les nouveaux cas de cancers[a] survenant dans leur population résidente; ils couvrent 22 départements, soit 20% de la population française.

- La production régulière des indicateurs de **surveillance épidémiologique du cancer en France** (incidence, mortalité, survie, prévalence) est assurée dans le cadre d'un partenariat regroupant:
  - Le réseau des registres français de cancer (Francim)
  - Le service de Biostatistique des HCL, qui héberge et analyse les données de la base commune des registres
  - Santé Publique France et l'Institut National du Cancer

- La production de ces indicateurs est très complète au niveau national, malgré la couverture partielle

- Au niveau **régional et départemental** en revanche, ces indicateurs sont plus difficiles à produire et cet axe a fait l'objet de nombreux travaux et développements méthodologiques ces dernières années

- Les **big-data**, utilisées en complément des données de registres, sont au cœur de ces travaux

[a] Recherche active multi-source, codage en CIM-O3 et selon les recommandations internationales

Problématique

⇒ **Comment estimer l'incidence départementale dans tous les départements**, alors que seuls 22 sont couverts par un registre ?

- Intérêt évident des proxy issus des big-data et notamment des Bases Médico-Administratives (BMA[b]), i.e. **données hospitalières et de l'assurance maladie (PMSI, ALD, SNIIR-AM…)**

- **Mais** un dénombrement direct des cas dans les BMA ne constitue pas une mesure correcte de l'incidence

- Afin d'estimer l'incidence départementale des cancers, une méthodologie appropriée est nécessaire
  Elle repose sur l'utilisation des **BMA en tant que « proxy»** de l'incidence, et non de mesure directe

Objectif

Une **méthodologie** a ainsi été progressivement développée [a] afin de :

- **Calibrer** le proxy sur l'incidence 'gold-standard' issue des registres,
  i.e. confronter le proxy à l'incidence et le corriger
- Calculer la **précision** des estimations en prenant en compte les « **erreurs de mesure** »
- **Evaluer la qualité** des estimations, pour chaque cancer
- Produire les estimations départementales, lorsqu'elles sont jugées de qualité suffisante

[a] Uhry 2007, Remontet 2008, Mitton 2011, Uhry 2013, Colonna 2013, Colonna 2015, Chatignoux 2017

Notations

Pour un cancer et un sexe donné :

a: âge (ou classe d'âge)

d département

$I_{a,d}$: nombre de cas incidents d'âge *a* enregistrés par le **registre** dans le dept *d*

$P_{a,d}$: nombre de nouveaux cas d'âge *a* dénombrés dans le **proxy** étudié dans le dept *d*,
comme par exemple:

- nombre de patients avec une 1ère hospitalisation avec en diagnostic principal le cancer étudié
- nombre de nouvelles demande d'ALD pour le cancer étudié

Méthodologie (1)

- **Modèle de calibration, zone registre**

  Dans les départements couvert par un registre, on modélise le rapport entre le proxy et l'incidence (P/I) selon l'âge *a :*

  $P/I_{a,d} = f(a)$

  Point technique : le modèle inclus aussi un effet dept aléatoire $N(0,\sigma^2)$ dans le cadre d'un modèle de Poisson à effets mixtes

  On obtient ainsi une estimation de ce rapport moyen dans la zone registre *f(a)*

  Par exemple, si pour un âge *a* donné, on dénombre 80 patients nouvellement hospitalisés pour 100 cas registres, alors *f(a)*=0,8 pour cet âge

- **Estimation de l' incidence dans un département sans registre, à partir du proxy**

  $I_a = P_a / f(a)$

  Par exemple, si **P=32** patients hospitalisés pour un âge a donné, et ***f(a)*=0,8** , alors le nombre de cas incidents pour cet âge est estimé par **I= 32 / 0,8=40** cas

Méthodologie (2)

- **Qualité des estimations et erreur de prédiction**

Point technique : prédictions en validation croisée

- Dans les depts avec registres, l'incidence est prédite avec le modèle de calibration et comparée à l'incidence observée: erreur de prédiction

- Lorsque les erreurs de prédictions sont jugées trop importantes, les estimations ne sont pas présentées

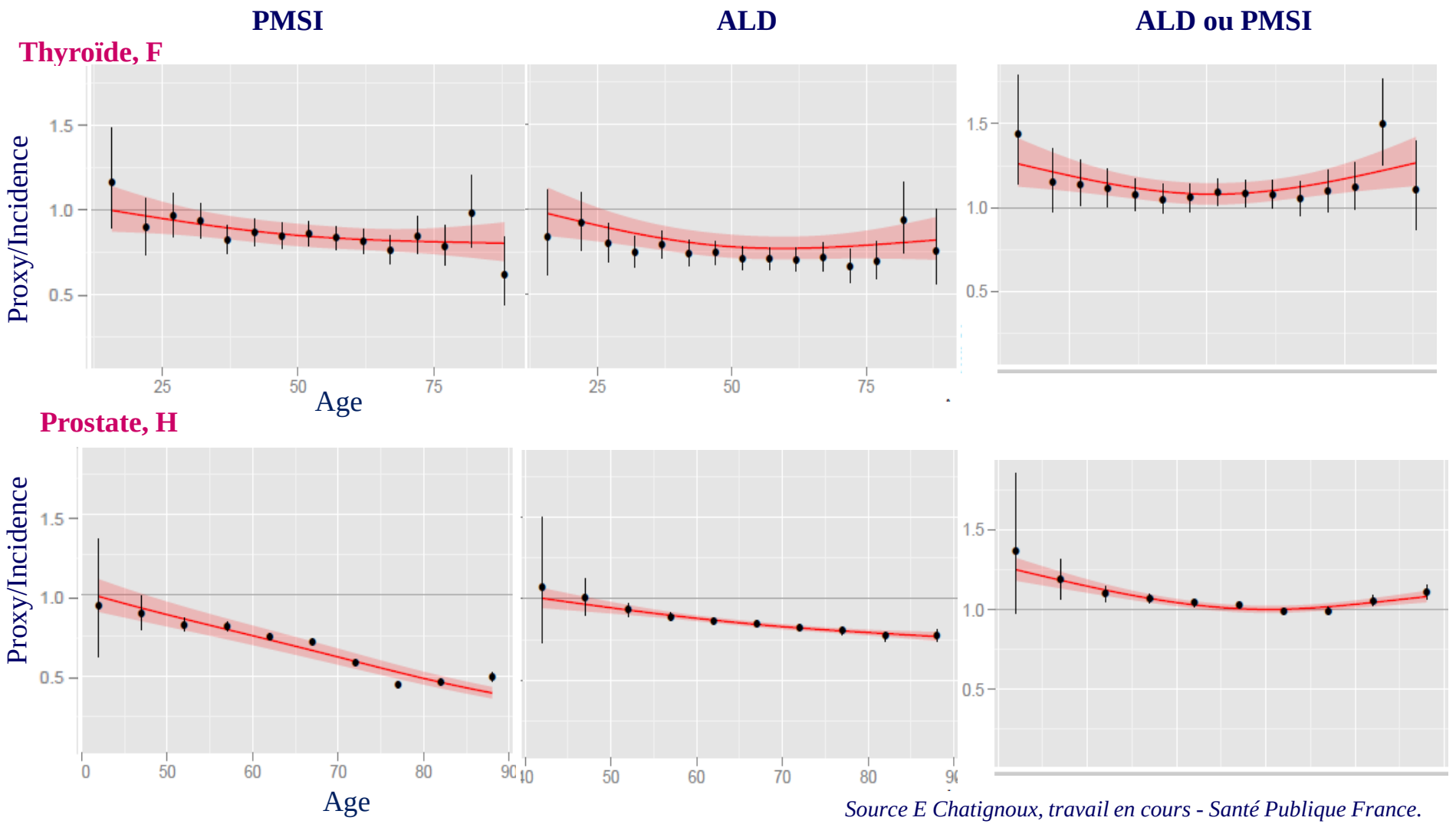
- A noter que la qualité de estimations (i.e. faibles erreurs) est directement liée au fait que le ratio P/I varie ou non selon le département

- **Précision des estimations**

- La précision des estimations prend en compte justement la variabilité départementale du ratio P/I

Point technique : intègre la variabilité deptale estimée σ issue du modèle de calibration

Illustration (1) — Calibration, rapport Proxy/Incidence selon l'âge (2007-2011)



Source E Chatignoux, travail en cours - Santé Publique France.

!!! Le dénombrement des cas dans les BMA n'est pas une mesure correcte de l'incidence !!!

… Et même lorsque ratio proche de 1, il y a des « faux positifs» et des faux «négatifs» …

Illustration (2)

Qualité des estimations / Erreurs de prédictions

ER: erreur de prédiction en % (Prédits/Obs) ; Proxy: ALD ou PMSI (AUP) , 2007-2011

Sein (Femmes), 2007-2011

Dépt.	Inc. obs.	AUP	Prédits	ER^d
14	2838	3141	2772.5	-2.3
21	2028	2200	1938.3	-4.4
25	2394	2591	2279.7	-4.8
33	4663	5390	4777.0	2.4
34	4316	5106	4526.9	4.9
38	4656	5146	4532.3	-2.7
44	5278	6260	5545.7	5.1
50	2031	2233	1970.0	-3.0
67	4305	4876	4307.4	0.1
68	2926	3312	2925.1	-0.0
80	2299	2602	2298.1	-0.0
81	1546	1775	1573.5	1.8
85	2705	3126	2763.8	2.2
87	1252	1390	1227.2	-2.0
TOT.	**43237**	**49148**	**43437.5**	-

=> BMA ne fournit pas une mesure correcte de l'incidence

=> Sein: ER (val. abs.) maximum de 5%

Foie, Hommes , 2007-2011

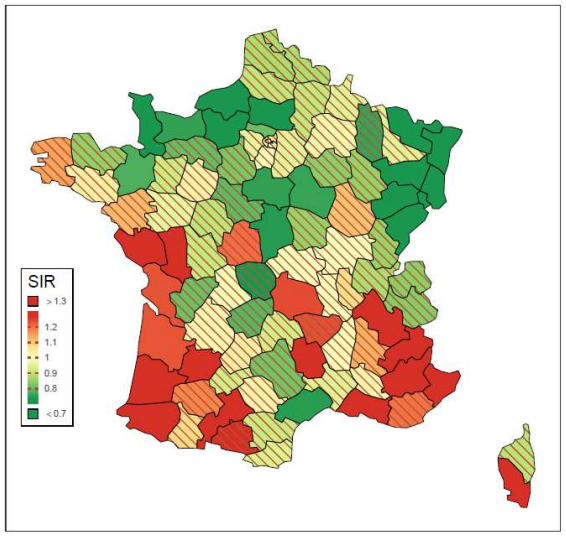
Dépt.	Inc. obs.	AUP	Prédits	ER^d
14	430	393	349.0	-18.8
21	313	331	296.3	-5.3
25	273	314	282.6	3.5
29	624	611	544.4	-12.8
33	509	559	500.8	-1.6
34	602	647	578.2	-4.0
38	623	546	483.4	-22.4
44	1056	1093	976.0	-7.6
50	294	312	279.2	-5.0
67	574	638	573.1	-0.1
68	330	414	375.3	13.7
71	295	416	381.3	29.3
80	258	279	250.0	-3.1
81	140	158	141.8	1.3
85	468	584	530.4	13.3
87	100	148	134.9	34.9
TOT.	**6889**	**7443**	**6676.7**	-

Source Chatignoux, travail en cours.

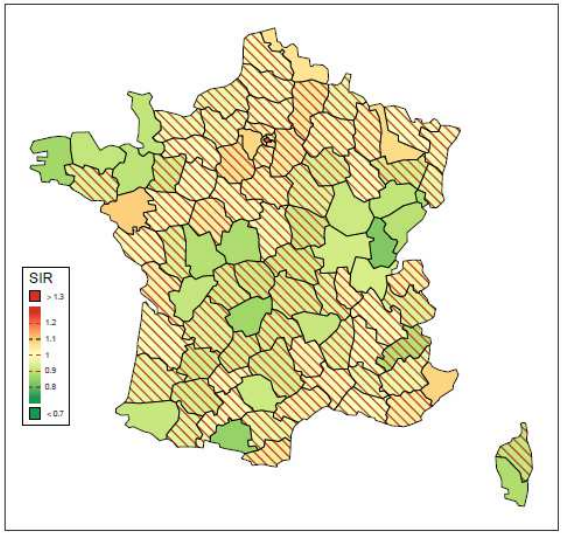
=> Foie: ER (val. abs.) de plus de 30% ..

Carte des SIR - Standardized Incidence Ratio[a]

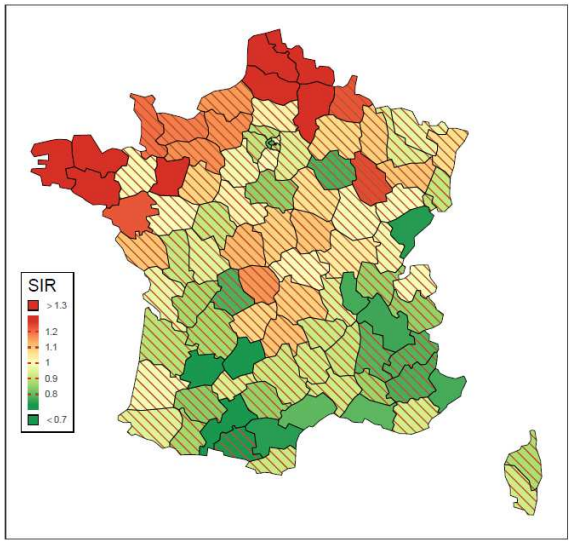
Thyroïde, Femmes



Sein



Oesophage, Hommes



Source Chatignoux, travail en cours et BEH 2016.

[a] Le SIR est une mesure d'écart à la moyenne nationale, avec une standardisation sur la structure d'âge.
SIR= (effectifs observés)/ (effectifs attendus si le dept avait les même taux pour chaque âge qu'au niveau national)

Conclusion et perspectives

- Les proxys issus de big-data, et en particulier des BMA, sont très utiles pour estimer l'incidence départementale des cancers

- En effet, dans la dernière évaluation réalisée (2007-2011, à paraître), sur la vingtaine de tumeurs solides étudiées, seuls 4 cancers présentaient des erreurs de prédictions jugées prohibitives

- Toutefois, les BMA ne **fournissent pas une mesure correcte de l'incidence** et donc **ne doivent pas être utilisées seules**

- Pour estimer l'incidence départementale des cancers, il faut une **méthodologie appropriée** :
  - **Calibrer** sur l'incidence des registres
  - Rendre compte des « **erreurs de mesure** » dans la précision des estimations
  - Evaluer la **qualité des estimations**

- Si les algorithmes de sélection peuvent être perfectionnés, cette méthodologie n'en reste pas moins nécessaire

- Le niveau infra-départemental reste un champ à ouvrir, qui pose de nouveaux défis méthodologiques..

Merci de votre attention ...

Références

Uhry Z, Colonna M, Remontet L et al. Estimating infranational and national thyroid cancer incidence in France from cancer registries data and national hospital discharge database. **Eur J Epidemiol. 2007**

Remontet L, Mitton N, Couris C et al. Is it possible to estimate the incidence of breast cancer from medico-administrative databases? **Eur J Epidemiol. 2008**

Mitton N, Colonna M, Trombert B et al. A suitable approach to estimate cancer incidence in area without cancer registry. **J Cancer Epidemiol. 2011**

Uhry Z, Remontet L, Colonna M et al. Cancer incidence estimation at a district level without a national registry: A validation study for 24 cancer sites using French health insurance and registry data. **Cancer Epidemiol. 2013**

Chatignoux E, Remontet L, Colonna M et al. For a sound use of big data in epidemiology: evaluation of a calibration model for count data with application to prediction of cancer incidence in areas without cancer registry. *Soumis*

Grosclaude P, Dentan C, Trétarre B et al. Etude des caractéristiques des bases de données médico-administratives permettant de les utiliser comme indicateurs de suivi épidémiologique des cancers. Comparaison avec les données des registres au niveau individuel. **Bull Epidémiol Hebd. 2012**

Colonna M, Mitton N, Remontet L et al. Méthode d'estimation de l'incidence régionale à partir des données d'incidence des registres, des données de mortalité et des bases de données médico-administratives. **Bull Epidémiol Hebd. 2013**

Colonna M, Mitton N, Remontet L et al. Incidence régionale des cancers 2008-2010. Évaluation de trois méthodes d'estimations: analyse et résultats. **Institut de veille sanitaire, 2014.** Rapport, 191 p.

Colonna M, Chatignoux E, Remontet L et al. Estimations de l'incidence départementale des cancers en France métropolitaine 2008-2010. Étude à partir des données des registres des cancers du réseau Francim et des bases de données médico-administratives. **Institut de veille sanitaire; 2015**. Rapport, 50 p.

Chatignoux L, Uhry Z, Remontet L et al. Estimations départementales de l'incidence du cancer de la thyroïde à partir des données des registres et du croisement de deux sources de données médico-administratives, France, 2007-2011. **Bull Epidémiol Hebd. 2016**