

Les Data Sciences pour la santé: Qui, Quoi, Comment ?

Pascal Staccini

IRIS Department - School of Medicine of Nice
Nice-Sophia Antipolis University
UMR 912 - SESSTIM - INSERM





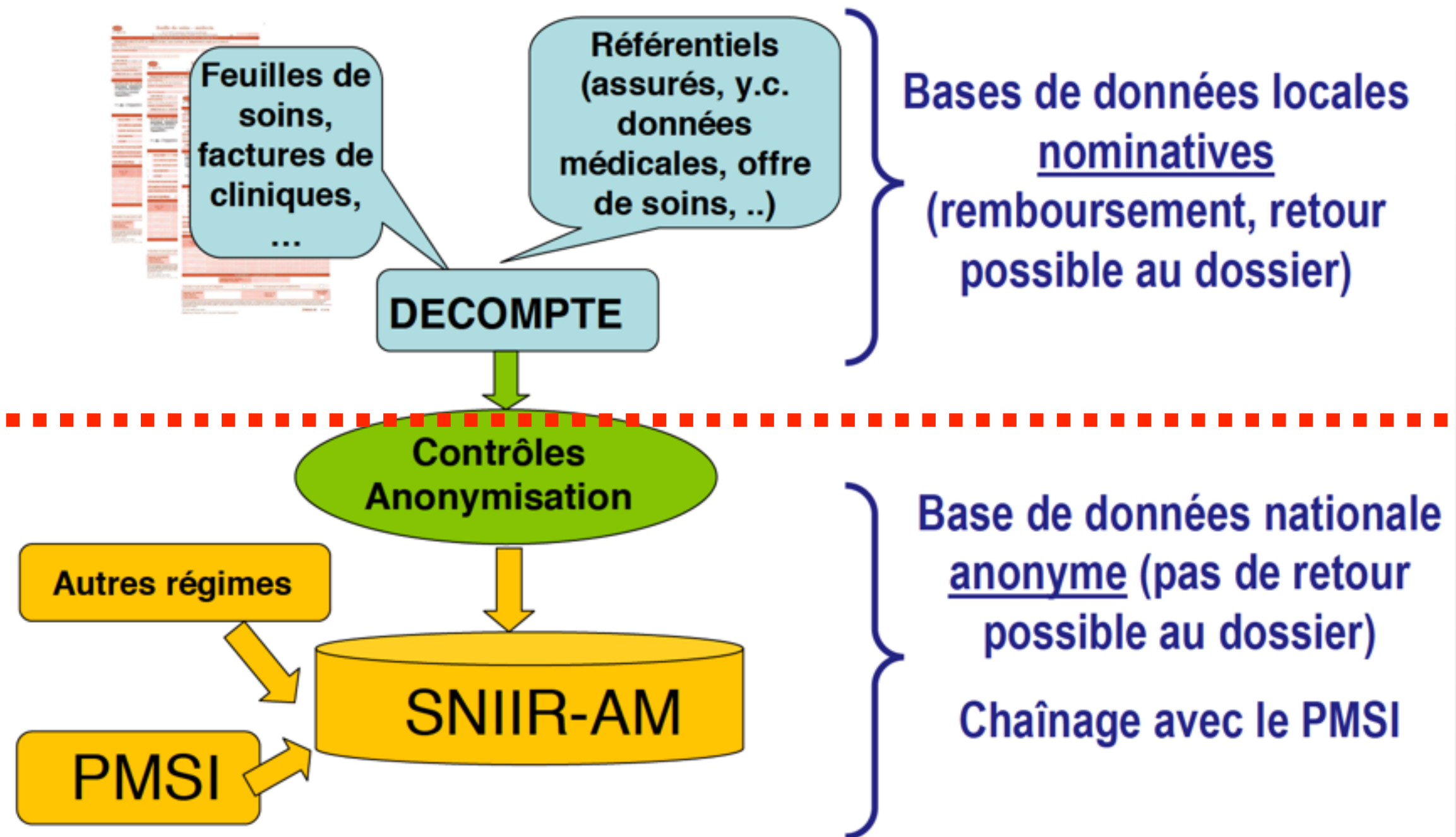
et pour quoi faire ?



Les données de santé

C'est quoi ?

Sophia Conf - 6 juillet 2015





Un système d'information centré recours

Données « Patient » structurées chaînables...
de façon ANONYME

Algorithme public IRREVERSIBLE

Sophia Conf - 6 juillet 2015



Quelles données ?

- Sur les **patients** :
 - Age, sexe, notion de CMU-C, diagnostic de l'ALD, département et région de résidence, date de décès
- Sur la **consommation de soins en ville** : (détail par date de soins et date de remboursement)
 - Toutes les prestations remboursées avec le codage détaillé de la prestation (actes médicaux avec le code CCAM, biologie, dispositifs médicaux, code CIP des médicaments)

Sophia Conf - 6 juillet 2015



Quelles données ?

- Sur la **consommation de soins en établissement** : (détail par date de soins et date de remboursement)
 - Les données du PMSI (MCO-SSR-HAD) pour l'ensemble des établissements sanitaires (motif médical, actes CCAM, durée de séjour, mois de sortie, mode de sortie,...)
 - L'activité externe des hôpitaux (depuis 2009)
 - Les médicaments et dispositifs facturés « en sus » des forfaits (GHS)



Quelles données ?

- Sur les **pathologies diagnostiquées / traitées / suivies**
 - Les codes CIM-10 pour les patients en affection de longue durée (ALD30) : 8 millions de personnes
 - Les codes CIM-10 issus du PMSI pour les séjours hospitaliers (diagnostics principaux, associés, reliés)
 - Des informations de nature médicale (GHM, médicaments traceurs, actes techniques réalisés par les professionnels de santé, examens biologiques ou dispositifs médicaux)

Sophia Conf - 6 juillet 2015



Limites

- **Pas de résultat** d'examen clinique ou paraclinique (tabagisme, niveau tensionnel, HbA1c, IMC,...)
- **Pas d'information** sur l'hébergement en structure **médico-sociale** des personnes âgées ou sur les hospitalisations en long séjour (en voie d'amélioration)
- **Pas d'information** médicale sur les séjours en centre hospitalier spécialisé (**psychiatrie**)
- **Pas, ou très peu, de données sociales** (uniquement notion de CMU-C)

Sophia Conf - 6 juillet 2015



From Small Data to « almost » Big Data

Sophia Conf - 6 juillet 2015



Big « amount of » data...

- Des milliards d'évènements traités : sur 30 mois 20 milliards de lignes de prestation
 - = 9,0 milliards de lignes de ventilation comptable
 - = 8.4 milliards de lignes de nature de prestation
 - = 3,4 milliards de lignes de pharmacie « affinée » (CIP)
 - = 1,2 milliard de lignes de biologie « affinée » (code biologie)
 - = 270 millions de lignes d'actes techniques médicaux (CCAM)
 - = 57 millions de lignes d'actes de transport médicaux

Sophia Conf - 6 juillet 2015

Quelles utilisations ?

- Données **observationnelles** : Données inférées (diagnostics) / Données objectives (actes, remboursements) / Données datées (au mois pour les recours hospitaliers) / Données géolocalisées (codes postaux « PMSI »)
- Analyse de **cohortes longitudinales** : 1) recherche d'une association statistique (incertitude d'échantillonnage) entre exposition et maladie : $P(M+) = f(\text{FDR1}, \text{FDR2}, \text{FDR3}, \dots)$, 2) quantification de cette association (risque relatif ou rapport de cotes) 3) qualification de la causalité (facteurs qualitatifs de temporalité, de dose-réponse, de physiopathologiques)

		%	HR brut	HR ajusté			%	HR brut	HR ajusté
Caractéristiques démographiques					Antécédents cardiovasculaires (6 mois avant l'hospitalisation index)				
Age en année (référence=65-74)	<=44	8,7	0,15***	0,17***	ALD cardiovasculaire		23,3	2,31***	1,45***
	45-54	18,5	0,32***	0,40***	Hospitalisation pour motif cardiovasculaire (<i>aucune</i>)	Stent	2,7	0,98	1,11
	55-64	19,8	0,54***	0,61***		Autre	8,5	2,46***	1,49***
	75-84	22,6	2,26***	1,61***	Hospitalisation index (et les 30 jours suivants)				
	>=85	9,7	5,14***	1,95***	Procédures thérapeutiques (<i>aucune</i>)	Pontage	1,8	0,31***	0,63*
Homme		68,7	0,49***	1,17**		Angioplastie	2,4	0,24**	0,53**
Bénéficiaire CMUC		5,9	0,54***	1,26		Stent	62,2	0,22***	0,56***
Cormorbidités °					Traitement médicamenteux post-infarctus (pendant 30 mois)				
Cancer		9,1	2,50***	1,97***	Statines (taux d'observance > 80%)	Non-consommant	14,3	11,23***	5,25***
Diabète		22,8	1,43***	1,30***		Tx obs <= 80%	20,8	2,47***	1,98***
Insuffisance rénale chronique		1,6	2,96***	1,98***	*p<0.05, **p<0.01, *** p<0.001				
Symptômes inflammatoires		2,7	1,84***	1,55***					
Maladies chroniques du foie		0,6	1,78*	2,00**					
					° Non significatif après ajustement : maladies respiratoires, maladie d'Alzheimer, maladie de Parkinson, dépression, affections psychiatriques				

Patients vivant après une hospitalisation pour infarctus du myocarde (régime général stricto sensu, 1er sem 2006) :
étude des facteurs associés au décès à 30 mois
 Sophia Conf - 6 juillet 2015



Causality

- **Adverse events using benfluorex**
- Cohort study exposed unexposed: 1 048 173 diabetic people aged from 40 to 69 including 43 044 exposed to benfluorex in 2006 (selection criteria of diabetic people: (≥ 3 reimbursed orders of oral anti-diabetic medication and/or insuline at different times))
- Search for specific events in DRG databases: hospitalizations for mitral or aortic insufficiency; surgery valve replacement for valvular insufficiency

Sophia Conf - 6 juillet 2015

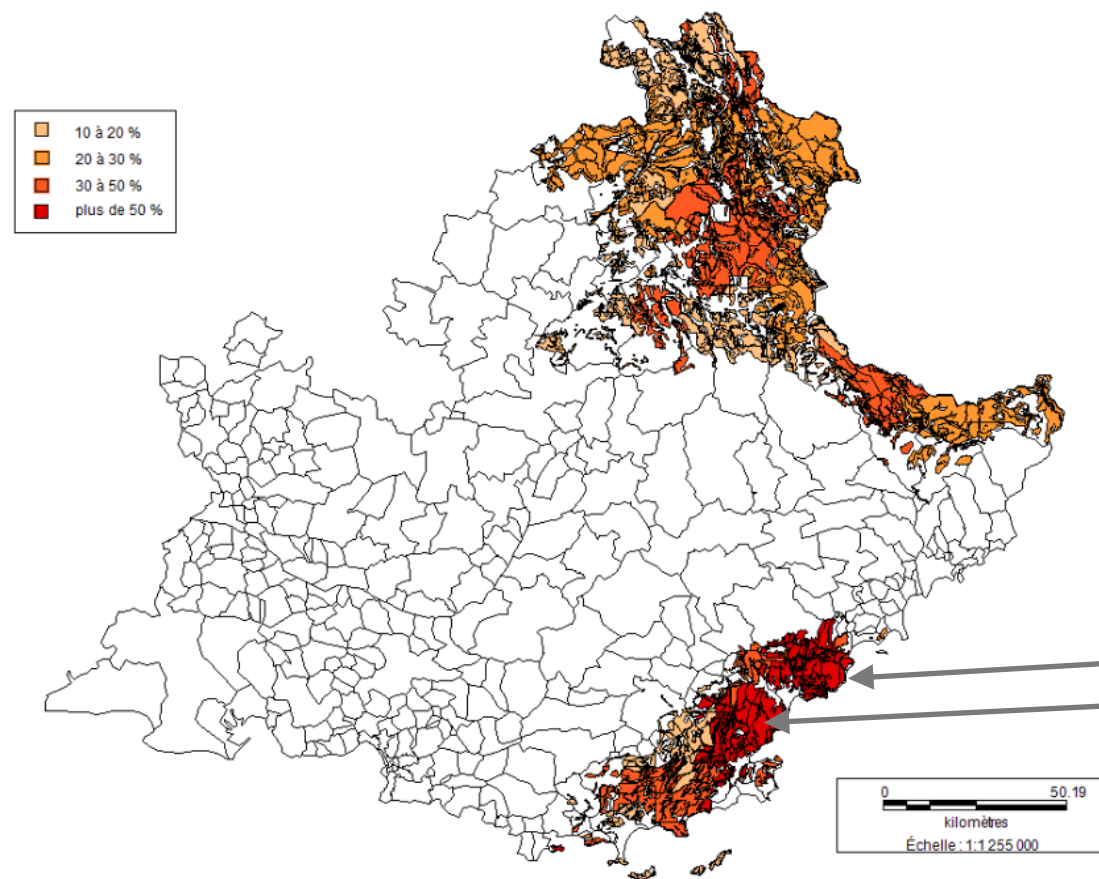
<i>Hospitalisation pour...</i>	Risque d'hospitalisation pour 100 000 personnes		RR brut (IC 95%)	RR ajusté ¹ (IC 95%)
	Non Exposées	Exposées		
Insuffisance valvulaire	27	76	2,9 [2,2 – 3,7]	3,1 [2,4 – 4,0]
Insuffisance mitrale	18	43	2,5 [1,8 – 3,5]	2,6 [1,9 – 3,7]
Insuffisance aortique	9	34	4,0 [2,7 – 5,9]	4,4 [3,0 – 6,6]
Rempl. valvulaire (IM-IA)	8	28	3,7 [2,4 – 5,6]	3,9 [2,6 – 6,1]

¹ ajustement sur l'âge, le sexe et la notion d'ALD cardiovasculaire

Significant **association**
and previous exposure

All hospital stays from 2006 to 2010 - PACA region Discharge data sets with a diagnosis of lung cancer (297 365)

Le pourcentage du bâti en Aléa 4 RADON

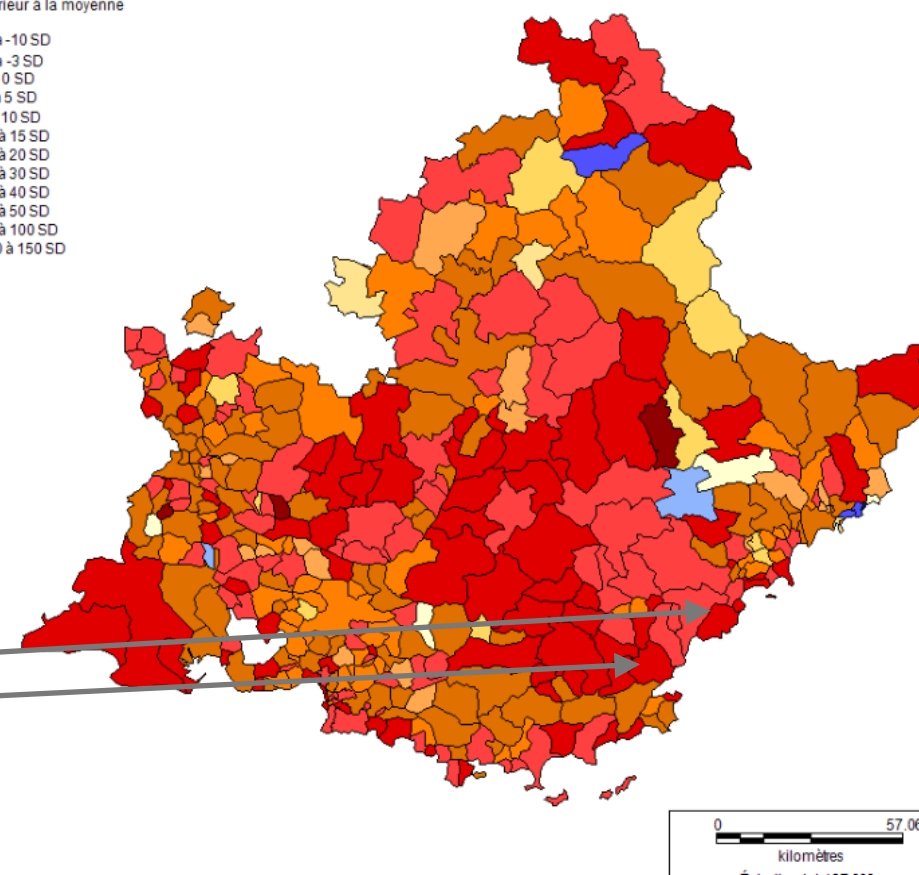


Séjours hospitaliers pour cancer du poumon : écarts à la moyenne nationale par PMSI

Jaune-orange-rouge: supérieur à la moyenne
Bleu: inférieur à la moyenne

Legend for hospital stays:

- 15 à -10 SD
- 10 à -3 SD
- 3 à 0 SD
- +0 à 5 SD
- +5 à 10 SD
- +10 à 15 SD
- +15 à 20 SD
- +20 à 30 SD
- +30 à 40 SD
- +40 à 50 SD
- +50 à 100 SD
- +100 à 150 SD



Geographical **correlation** Radon exposure and lung cancer



Cas concret de « profilage »

Utilisation des matrices de co-occurrence

Sophia Conf - 6 juillet 2015

Cas concret 'PMSI'

- **3 positions du « diagnostic » dans les résumés de sortie :**
 - Diagnostic principal (défini rétrospectivement, qui a motivé le recours) / Diagnostic relié (lorsque DP en Z – bilan, chimio- radio- thérapie) / Diagnostic associé (complications, associations, antécédents...)



Cas concret 'PMSI'

- Ces associations correspondant à **3 circonstances** :
 - l'acte pratiqué permet de faire le **diagnostic** (positif, différentiel) ou contribue à celui-ci ;
 - l'acte pratiqué entre dans le cadre du **bilan** (extension, gravité) de la pathologie du patient ;
 - l'acte pratiqué permet de **traiter** la pathologie représentée par le diagnostic codé.

Sophia Conf - 6 juillet 2015



Cas concret 'PMSI'

- Les données sources utilisées sont les 798 916 RSA 2009 de la région PACA, à l'exclusion des RSA correspondant à des séances répétitives ou à des hospitalisations pour bilan (diagnostic principal en Z, maladie en DR).
- 299 715 cooccurrences diagnostic principal – acte ont été étudiées (celles dont l'acte n'avait qu'une signification tarifaire ont été exclues).
- Travail mené conjointement avec Rouen (SJ Darmoni)

Sophia Conf - 6 juillet 2015

Cas concret 'PMSI'

- Pour chaque cooccurrence, la force de la cooccurrence a été mesurée à l'aide du coefficient de **confiance DP,Acte**
Il s'agit de la probabilité d'observer l'acte dans un résumé de sortie en présence du DP (diagnostic principal) :
- $\text{confiance DP-Acte} = P(\text{Acte}/\text{DP})$

Cas concret 'PMSI'

- **La confiance Acte,DP** qui est la probabilité d'observer le DP dans un résumé en présence de l'acte :
- $\text{confiance Acte-DP} = P(\text{DP}/\text{Acte})$
- **Le lift** qui rapporte la confiance à la probabilité de l'élément co-occurent :
- $\text{lift} = \text{confiance DP,Acte} / P(\text{Acte})$

Cas concret 'PMSI'

- Sélection des cooccurrences dont le diagnostic principal correspond à un diagnostic de la famille « cancer colique » (codes CIM10 débutants par C18).
- Filtrage des cooccurrences par les métatermes CISMeF « gastroentérologie » et « chirurgie digestive ».

Sophia Conf - 6 juillet 2015

Cas concret 'PMSI'

LIB_MT	DiagCancerColon	CodActe	libelle_acte	NbCooc	NbKColique	NbActesTotal	Confiance Diag	Confiance Acte	Lift	Coef vraisemblance
chirurgie générale	cancer colique	HHFA009	Colectomie droite avec rétablissement de la continuité, par laparotomie	652	3490	1105	0,1868	0,5900	135,0707	165,9
chirurgie générale	cancer colique	HHFA022	Colectomie totale avec conservation du rectum, avec anastomose iléorectale, par laparotomie	43	3490	73	0,0123	0,5890	134,8408	136,5
chirurgie générale	cancer colique	HHFA018	Colectomie transverse, par laparotomie	159	3490	328	0,0456	0,4848	110,9683	116,2
chirurgie générale	cancer colique	HHFA006	Colectomie gauche avec libération de l'angle colique gauche, avec rétablissement de la continuité, par laparotomie	261	3490	569	0,0748	0,4587	105,0035	113,4
chirurgie générale	cancer colique	HHFA008	Colectomie droite avec rétablissement de la continuité, par coelioscopie ou par laparotomie avec préparation par coelioscopie	233	3490	529	0,0668	0,4405	100,8268	108,0
chirurgie générale	cancer colique	HHFA023	Colectomie transverse, par coelioscopie ou par laparotomie avec préparation par coelioscopie	38	3490	82	0,0109	0,4634	106,0829	107,2
chirurgie générale	cancer colique	HHFA017	Colectomie gauche sans libération de l'angle colique gauche, avec rétablissement de la continuité, par laparotomie	59	3490	162	0,0169	0,3642	83,3706	84,8
chirurgie générale	cancer colique	HHFA002	Colectomie gauche avec libération de l'angle colique gauche, avec rétablissement de la continuité, par coelioscopie ou par laparotomie avec préparation par coelioscopie	206	3490	837	0,0590	0,2461	56,3401	59,8
chirurgie générale	cancer colique	HHFA010	Colectomie gauche sans libération de l'angle colique gauche, avec rétablissement de la continuité, par coelioscopie ou par laparotomie avec préparation par coelioscopie	28	3490	119	0,0080	0,2353	53,8625	54,3
chirurgie générale	cancer colique	HHFA026	Colectomie droite sans rétablissement de la continuité, par laparotomie	27	3490	131	0,0077	0,2061	47,1811	47,5
chirurgie générale	cancer colique	HJFA002	Résection rectosigmoïdienne avec anastomose colorectale infrapéritonéale, par laparotomie	95	3490	546	0,0272	0,1740	39,8297	40,9
chirurgie générale	cancer colique	HLFA019	Résection atypique du foie, par laparotomie	38	3490	244	0,0109	0,1557	35,6508	36,0
chirurgie générale	cancer colique	HGFA007	Résection segmentaire unique de l'intestin grêle avec rétablissement de la continuité, en dehors de l'occlusion, par laparotomie	79	3490	532	0,0226	0,1485	33,9931	34,8
chirurgie générale	cancer colique	HGCA004	Entérostomie ou colostomie cutanée de protection, au cours d'une résection intestinale avec rétablissement de la continuité	23	3490	157	0,0066	0,1465	33,5354	33,8
chirurgie générale	cancer colique	HGCA008	Entérostomie cutanée, par laparotomie	33	3490	268	0,0095	0,1231	28,1874	28,4
chirurgie générale	cancer colique	HHFA014	Colectomie gauche sans libération de l'angle colique gauche, sans rétablissement de la continuité, par laparotomie	29	3490	240	0,0083	0,1208	27,6607	27,9
chirurgie générale	cancer colique	HHCA002	Colostomie cutanée, par laparotomie	68	3490	582	0,0195	0,1168	26,7462	27,3
chirurgie générale	cancer colique	HJFA004	Résection rectosigmoïdienne avec anastomose colorectale infrapéritonéale, par coelioscopie ou par laparotomie avec préparation par coelioscopie	66	3490	639	0,0189	0,1033	23,6439	24,1
chirurgie générale	cancer colique	HMFA007	Cholécystectomie, par laparotomie	57	3490	1048	0,0163	0,0544	12,4506	12,6
chirurgie générale	cancer colique	ZCQA001	Exploration de la cavité abdominale, par laparotomie [Laparotomie exploratrice]	28	3490	679	0,0080	0,0412	9,4398	9,5



Cas concret 'PMSI'

- Intérêt des co-occurences : Retrouver les pratiques médicales à partir des pratiques de codage ? Déterminer des variations - évolution des pratiques ? Améliorer la qualité des données (erreurs de codage)

Cas concret 'PMSI'

- Utilisation de filtres sémantiques (méta termes MeSH développés par Rouen) qui identifient le contexte de la présence de la cooccurrence. Il enlève du bruit et renforce donc la « pertinence » du lien-cooccurrence diag-acte
- De façon non spécifique (approche tous diagnostics)
- De façon spécifique (diagnostics spécifiés / exemple cancer)

Sophia Conf - 6 juillet 2015



Data Sciences and Healthcare Data

- exhaustivité du reporting...
- temporalité (chronologie des événements)
- géolocalisation (taille maillage corrélation environnemental)
- choix des seuils de discrimination (suspicion, état avéré)
- sémantique des classifications (et liens inter-classifications)
- chaînage anonymisé des sources connues (médico-économiques) et à venir ((physiologiques et d'auto-mesure, médicales, génotypiques, sociales, économiques)

Sophia Conf - 6 juillet 2015



Data Sciences and Healthcare Data

- Travail sur site pour qualifier les données (complétude et erreurs)
- Approximation du niveau et de la durée d'exposition
- Temps réel (ou délai raisonnable) entre la captation et le traitement des données (tendances mensuelles...)

Sophia Conf - 6 juillet 2015



Formation académique

- Spécialité de master 2 « Ingénierie pour les Systèmes de Santé et l'Autonomie » (I2SA) - Collaboration IRIS-I3S - En cours de définition :
 - données d'observation, données d'intervention, données médico-économiques, données de surveillance des process santé
 - data management : sources, chaînage / anonymisation, méthodes et outils qualité, ETL, calage temps / localisation entre les sources, back office des ontologies
 - inference management multi-sources : mesures d'association probabilistes / statistiques / sémantiques

Sophia Conf - 6 juillet 2015



Big Conclusion

- Big data for **better understanding** « good health » / « wellness »
- Big data for better understanding the **context** / the environment and the **interactions** with individuals
- Big amount of data **extracted** from health services databases
 - **Retrospective** epidemiological studies: identification of **risk factors**, phenotypes / contrast with « small data »
- Big amount of data **prospectively** recorded by means of sensors
 - **Real time** correlation between environmental conditions and patient's health profile / status



Big Data for healthcare?

P⁴P

From **Patient** for **Patients**

From **Physicians** for **Patients**

From **Physicians** to **Physicians**

Practice for **Performance**

Practice for **People**



Merci de votre attention

Pour toute réclamation :
pascal.staccini@unice.fr

Sophia Conf - 6 juillet 2015