Modélisation statistique des erreurs dans le couplage d'enregistrements appliquée aux données du Registre du cancer du SEER

Michael D. Larsen (GWU)
Collaborateurs :
Will Howe, Nicola Schussler (IMS),
Benmei Liu, Valentina Petkov, Mandi Yu (NCI)

Symposium international de 2016 sur les questions de méthodologie de Statistique Canada
Le mercredi 23 mars, 10 h 30 à 12 h 00
Séance 6A

Aperçu

1. Cas de cancer du sein du SEER et test Oncotype DX de GHI
   - GHI=*Genomic Health Incorporated*
2. Couplage d'enregistrements
3. Plan de l'examen manuel
4. Résultats
5. Conclusions

Avertissement : Les opinions formulées dans la présente communication sont celles de l'auteur et ne représentent pas nécessairement celles d'une autre personne ou organisation.

Registres du cancer du sein du SEER

- Programme *Surveillance, Epidemiology and End Results* (SEER) du *National Cancer Institute* (NCI)
- Registres du cancer fondés sur la population; 30 % des États-Unis; plusieurs secteurs.
- Données démographiques sur le patient, siège primaire de la tumeur, morphologie de la tumeur et stade au moment du diagnostic, première série de traitements et suivi du statut vital
- Objectifs et détails en ligne à **seer.cancer.gov**

Test OncoType DX de Genomic Health

- **Oncotype DX a été élaboré par Genomic Health, Inc. (GHI) en 2004**.
- Indiqué dans le cas du cancer du sein à un stade précoce (récepteurs hormonaux positifs, résultats négatifs au sujet des ganglions lymphatiques), afin de **stratifier le risque** de récurrence à long terme et d'aider à prédire les avantages de la chimiothérapie conjuguée à l'hormonothérapie.
- La qualité et l'intégralité des données peuvent être grandement améliorées si les renseignements sont obtenus directement des laboratoires qui procèdent au test moléculaire/génomique.
- **GHI est le seul laboratoire aux États-Unis qui effectue le test OncoType DX, ce qui en fait une cible idéale pour vérifier les couplages des résultats de laboratoire et des données du SEER.**

Couplage d'enregistrements du SEER et des fichiers de GHI dans des secteurs de registre

- Identifier les paires d'enregistrements qui concernent la même personne; combiner l'information des deux sources pour de vrais appariements.
- Transformer les comparaisons des variables en un ***score*** de similitude
- Scores élevés = appariement probable
  Faible score = non-appariement probable
- Des erreurs sont commises en raison d'erreurs dans les données, de valeurs manquantes et du caractère non unique
- Approche intermédiaire : une revue manuelle est possible

Logiciel LinkPlus 3.0 Beta

- Programme de couplage d'enregistrements probabiliste élaboré à la *Division of Cancer Prevention and Control* du CDC, à l'appui du *National Program of Cancer Registries* (NPCR) du CDC (gratuit en ligne)
- Fondé sur Fellegi et Sunter (JASA,1969)

Procédure de couplage globale

- Appariement en deux étapes
  - Première étape : LinkPlus pour obtenir les scores
  - Deuxième étape : programme SAS élaboré à l'interne pour finaliser les appariements
  - Nous avons expérimenté avec quelques seuils LinkPlus pour établir un équilibre entre la sensibilité menant au rejet des vrais appariements et les efforts de revue manuelle nécessaires (ainsi que la MÉMOIRE)
  - Élimination de doublons du SEER au niveau du patient; appariement à des cas du GHI; une fois les paires d'enregistrements déterminées comme correspondant à la même personne, associer les enregistrements des deux ensembles de données

Réglages de LinkPlus : Variables pour les pochettes

Variables pour les pochettes: Si les enregistrements correspondent exactement sur L'UN OU L'AUTRE de ces champs, on attribuera un score à l'appariement (assez large)

- État
- Prénom (Soundex)
- Nom de famille (Soundex)
- SSN
- Date de naissance

Réglages de LinkPlus : Variables d'appariement

Variables d'appariement : Utilisées pour les calculs des scores. Les appariements exacts obtiennent un score plus élevé que les appariements partiels. Les algorithmes d'appariements exacts se trouvent dans la boîte noire de LinkPlus. Pour chaque enregistrement du fichier principal, seul l'appariement comportant le meilleur score sera conservé :

- Prénom
- Deuxième prénom
- Nom de famille
- SSN
- Date de naissance

Méthodes : exigences additionnelles pour l'acceptation du couplage

Élaboration à l'interne fondée sur SAS (par IMS)

Parmi les paires dont le score est supérieur à 7 :

Appariement = appariement exact sur le prénom et le nom de famille et au moins deux des éléments suivants : date de naissance, SSN, (numéro de téléphone ou adresse)

Revue manuelle = critère intermédiaire

Non-appariement = non-appariement exact ou partiel sur trois des éléments suivants : prénom, nom de famille, date de naissance, SSN, numéro de téléphone, adresse* (ville et état)

* L'adresse n'est pas vérifiée dans le cas des appariements partiels

Questions de recherche

1. Quelle est l'exactitude du couplage?
2. Qu'est-ce qui affecte la qualité du couplage?

Étude d’évaluation

Plan de revue manuelle : examen de base

- Passer en revue les 18 643 appariements possibles dont le score est supérieur à 7 et qui sont classés comme « consultation manuelle »

Tous les enregistrements
CT: n=18 792 paires au-dessus du seuil de 7

Registry	SEER Records Provided Prior to De-duplication	SEER Records Provided: De-duplicated (reference)	Forward Linkage LinkPlus results Cutoff (lower limit) set to 7				Reverse Linkage LinkPlus results Cutoff (lower limit) set to 7			
			Best Match Score Below the Cutoff		Best Match Score Above the Cutoff		Best Match Score Below the Cutoff		Best Match Score Above the Cutoff	
	N	N	N	Pct (Reg)	N	Pct (Reg)	N	Pct (GHI)	N	Pct (GHI)
CA-Total^	261,015	248,151	151,041	60.9	97,110	39.1	278,523	89.2	133,062	32.3
CA-GCA	142,650	135,673	82,097	60.5	53,576	39.5	343,327	83.4	68,258	16.6
CA-LAX	64,561	61,305	37,208	60.7	24,097	39.3	377,225	91.7	34,360	8.3
CA-SFSJ	53,804	51,173	31,736	62.0	19,437	38.0	381,141	92.6	30,444	7.4
CT	35,955	33,621	20,748	61.7	12,873	38.3	392,793	95.4	18,792	4.6
GA	68,052	65,131	36,976	56.8	28,155	43.2	373,809	90.8	37,776	9.2
HI	11,290	10,793	7,433	68.9	3,360	31.1	406,524	98.8	5,061	1.2
IA	24,781	23,677	15,518	65.5	8,159	34.5	400,587	97.3	10,998	2.7
KY	33,268	31,879	18,585	58.3	13,294	41.7	394,751	95.9	16,834	4.1
LA	32,772	31,321	18,940	60.5	12,381	39.5	395,369	96.1	16,216	3.9
MI-DT	34,804	32,984	20,078	60.9	12,906	39.1	392,699	95.4	18,886	4.6
NJ	80,435	75,780	45,648	60.2	30,132	39.8	368,402	89.5	43,183	10.5
NM	13,574	12,909	7,105	55.0	5,804	45.0	404,137	98.2	7,448	1.8
UT	13,549	12,916	7,974	61.7	4,942	38.3	404,335	98.2	7,250	1.8
WA-SE	39,816	37,452	23,651	63.2	13,801	36.8	390,778	94.9	20,807	5.1
GHI#		411,585	n/a	n/a	n/a	n/a	n/a	n/a	336,313	81.7

Meilleurs appariements : Traitement au moyen de la méthode 4.1 CT: n=743 examens de revue manuelle

Registry	SAS Match Status Results for Method 4.1								
	Match			Manual Review			Non-match		
	N	Pct (AC)	Pct (GHI)	N	Pct (AC)	Pct (GHI)	N	Pct (AC)	Pct (GHI)
CA-Total	21,606	16.2	5.2	9,963	7.5	2.4	101,493	76.3	24.7
CA-GCA	11,784	17.3	2.9	5,562	8.1	1.4	50,912	74.6	12.4
CA-LAX	4,627	13.5	1.1	2,929	8.5	0.7	26,804	78.0	6.5
CA-SFSJ	5,195	17.1	1.3	1,472	4.8	0.4	23,777	78.1	5.8
CT	4,464	23.8	1.1	743	4.0	0.2	13,585	72.3	3.3
GA	9,666	25.6	2.3	1,511	4.0	0.4	26,599	70.4	6.5
HI	1,119	22.1	0.3	617	12.2	0.1	3,325	65.7	0.8
IA	2,456	22.3	0.6	366	3.3	0.1	8,176	74.3	2.0
KY	3,727	22.1	0.9	611	3.6	0.1	12,496	74.2	3.0
LA	4,016	24.8	1.0	549	3.4	0.1	11,651	71.8	2.8
MI-DT	4,729	25.0	1.1	774	4.1	0.2	13,383	70.9	3.3
NJ	11,367	26.3	2.8	1,914	4.4	0.5	29,902	69.2	7.3
NM	2,242	30.1	0.5	503	6.8	0.1	4,703	63.1	1.1
UT	1,459	20.1	0.4	253	3.5	0.1	5,538	76.4	1.3
WA-SE	4,118	19.8	1.0	839	4.0	0.2	15,850	76.2	3.9
Total	70,969	21.1	17.2	18,643	5.5	4.5	246,701	73.4	59.9

Plan de revue manuelle: paires additionnelles

1. Tirer un échantillon d'enregistrements qui obtiennent un score de 6 ou 7
2. Tirer un échantillon d'enregistrements qui obtiennent un score supérieur à 7 et qui sont des « appariements » selon les critères additionnels
3. Tirer un échantillon d'enregistrements qui obtiennent un score supérieur à 7 et qui sont des « non-appariements » selon les critères additionnels
4. Aussi OncoType=OUI dans SEER, mais non apparié (*n*=103)

Résultats : Nombre de paires appariées

		n	Liés	Non liés
1	Score de 5 à 6	1 999	0	1 999 (100 %)
2	Score > 7, appariement désigné	1 998	1 998 (100 %)	0
3	Score > 7, non-appariement désigné	1 998	0	1 998 (100 %)
4	Score > 7, groupe pour la revue manuelle	18 644	12 783 (70 %)	5 661 (30 %)
	Total	24 742	14 781 (60 %)	9 858 (40 %)

Groupes 1, 2 et 3: échantillons proportionnels par registre

Groupe 4: N = taille de la population de toutes les paires d'enregistrements

Conclusion : le score de 7 constitue un bon seuil.

Les critères additionnels d'appariement et de non-appariement sont précis.

La revue manuelle est assez importante.

Selon le résultat de la validation pour le SEER, OncotypeDX = Oui dans 4 registres

- 103 cas de la C.-B. avec OncotypeDX = Oui n'avaient pas d'appariement – manque de variables d'appariement
- 680 cas de la C.-B. avec OncotypeDX = Oui et appariement possible ont été rejetés suite à la revue manuelle – encore une fois, manque de variables d'appariement

Au total, 2 112 cas de la C.-B. avec OncotypeDX = Oui n'ont pas été appariés aux tests de GHI : 8,3 % du total des tests OncotypeDX (variation aussi selon le registre)

Étude des variables employées pour le couplage

Plusieurs variables ont été créées au moyen de SAS à l'interne pour les paires LinkPlus

- Ville, État, Rue : non-appariement, appariement, manquant [3]
- JJ, MM, AAAA : 3 versions + mineur + transposition [5]
- SSN, numéro de téléphone : 5 versions + JW [6]
- Nom de famille : 6 versions + contenu [7]
- Date de naissance : 6 versions + MD_swap [*non utilisé ici*]
- Deuxième prénom : 7 versions + 2 comparaisons avec le nom de famille [9]
- Prénom : 9 versions + 2 comparaisons avec le deuxième prénom [11]
  - *La distance de Jaro-Winkler n'est pas utilisée ici*

Score de prédiction

- Le R-carré pour le score de prédiction à partir des principaux effets de 10 variables est à 73 %.
- Toutes les variables comportent 2 niveaux ou plus statistiquement significatifs pour prédire le score.
- Répercussions sur le score si une paire comporte un non-appariement sur…

État	**-0,49**		**Deuxième prénom**	**-0,14**
SSN	**-0,20**		**Numéro de téléphone**	**-0,12**
Nom de famille	**-0,20**		**Prénom**	**-0,06**
Année	**-0,19**		**Nom de la rue**	**-0,05**
Jour	**-0,17**		**Mois**	**-0,04**

Prédiction de l'appariement au moyen de la régression logistique

- L'exactitude pour la prédiction de l'appariement (au moyen d'une probabilité estimée supérieure à 0,6) est de 92 %.
- Toutes les variables comportent 2 niveaux statistiquement significatifs ou plus pour prédire l'appariement.
- Répercussions du non-appariement sur une échelle linéaire…

SSN	**-5,86**		**Nom de la rue**	**-2,63**
Année	**-4,34**		**Mois**	**-2,61**
Nom de famille	**-3,55**		**État**	**-1,94**
Jour	**-3,50**		**Prénom**	**-1,76**
Numéro de téléphone	**-2,74**		**Deuxième prénom**	**-1,00**

Limites

- LinkPlus a été réglé pour produire uniquement le meilleur appariement
  - Un deuxième ou un troisième enregistrement peut représenter un quasi-appariement et ainsi aider à décider d'accepter ou non le meilleur appariement
- Vous devez procéder à votre propre comparaison de champs séparément pour intégrer ces données
- L'examen des enregistrements ne s'est pas fait à l'aveugle – les réviseurs savaient quels enregistrements de lot étaient présents et connaissaient le score de couplage – ceci est difficile à éviter

Sommaire

- Le couplage d'enregistrements a permis de déterminer efficacement la plupart des paires entre les cas de cancer du sein du SEER et la base de données Oncotype DX de GHI.
- LinkPlus comporte certaines limites qui ont été notées :
  - Limité à 10 variables d'appariement
  - Limite de mémoire
- La variabilité selon le registre du SEER sera étudiée
- La qualité des variables et la façon dont elles sont traitées de façon préliminaire sont considérées comme des facteurs clés de la réussite du couplage d'enregistrements.
- Certains résultats intéressants notés au chapitre du score de prédiction et de l'appariement, mais on doit poursuivre les travaux.

Avenir

- Travaux en cours pour établir des normes de rendement et de rapport pour les projets de couplage d'enregistrements du NCI.
- Comparaison d'autres logiciels de couplage d'enregistrements et méthodes de traitement des accords inexacts sur les champs d'information.
- Efforts de modélisation – quelles sont les répercussions du couplage d'enregistrements sur les analyses subséquentes.

Merci!

- Merci aux organisateurs et au FCSM, ainsi qu'au président et au commentateur de cette séance.
- Merci à mes co-auteurs et collaborateurs (NCI, IMS)
- Financement en vertu d'un contrat avec le NCI
- ***Merci à tous ceux qui ont procédé à la revue manuelle dans plusieurs bureaux de registre du SEER!***

mlarsen@bsc.gwu.edu