

Evaluation de la valeur discriminante des traits d'identification pour le chaînage des informations

Quantin Catherine^a, Binquet Christine^a, Bourquard Karima^b,
Allaert François-André^c, Pattisina Ronny^a, Gouyon-Cornet Béatrice^d, Ferdynus
Cyril^d, Gouyon Jean-Bernard^d.

^a Département d'information médicale, CHU de Dijon

^b Groupe de travail sur la Modernisation du Système d'Information Hospitalier (GMSIH)

^c W2 "data security" European Federation of Medical Informatics

^d Service de Pédiatrie, CHU de Dijon, France

INTRODUCTION

Si l'on veut reconstituer l'histoire médicale d'un patient, il est souvent nécessaire de combiner des informations provenant de différentes sources. En effet, il faut dans ce cas colliger toutes les informations disponibles concernant les recours antérieurs au système de soins (consultations, hospitalisations,...). Dans ce cas, il est indispensable d'éviter toute affectation erronée d'antécédents médicaux d'un autre patient qui pourraient conduire à une décision médicale inadaptée. Grâce au développement d'équipements informatiques de plus en plus performants, la corrélation d'informations relatives à même patient utilisant des méthodes probabilistes s'est largement répandue aussi bien dans le cadre d'études épidémiologiques que pour la prise en charge des patients. Cependant, même si ce type de méthode permet de limiter l'impact, par exemple, d'es erreurs de saisie, sur la qualité de la corrélation, plusieurs travaux méthodologiques (1-4) ont montré que la qualité de celle-ci dépend du pouvoir discriminant des traits d'identification utilisés. La mise en œuvre d'une corrélation avec des traits d'identification peu informatifs, pouvant se traduire par des erreurs de corrélation, par excès ou par défaut, il est donc essentiel de quantifier l'information apportée par chaque trait d'identification.

OBJECTIF

L'objectif de ce travail a été d'estimer les valeurs discriminantes des traits d'identification susceptibles de pouvoir être utilisés pour une corrélation. Cette estimation a été réalisée pour chaque trait séparément, mais aussi pour différentes combinaisons de ces traits. En effet, deux traits de valeur discriminante individuelle très élevée, peuvent une fois combinés, ne pas apporter autant d'information que l'on souhaiterait dans la mesure où ces traits sont redondants. Il convenait donc de calculer le poids combiné de ces traits.

MATERIEL ET METHODE

Matériel

Traits d'identification candidats

Six traits d'identification ont été envisagés : 1) Date de naissance ; 2) Nom de jeune fille (NJF) ; 3) Nom usuel (N) : en cas de nom usuel non renseigné, le NJF a été utilisé ; 4) Premier prénom ; 5) Second prénom ; 6) Sexe. Du fait de la fréquence des erreurs de saisie des noms et prénoms, nous avons choisi d'évaluer l'intérêt d'appliquer un traitement phonétique avant corrélation pour limiter leur impact. Deux types de traitements phonétiques ont été testés : le Soundex et un traitement phonétique adapté à la langue française (5, 6). Un traitement reposant sur la première lettre du prénom n'a pas été envisagé dans la mesure où le Soundex est déjà composé de la première lettre du prénom plus une numérotation dépendant des lettres suivantes. Par ailleurs, les noms usuels et noms de jeunes filles sont souvent intervertis. Nous avons tenu compte de cette inversion en utilisant un ou inclusif. Si un premier enregistrement contenait le nom usuel du sujet (N1) dans le champ 1 et le nom de jeune fille (NJF1) dans le champ 2 et qu'un second

enregistrement correspondant au même sujet contenait le nom de jeune fille (N2) dans le champ 1 et le nom usuel (NJF2) dans le champ 2, nous avons considéré que ces enregistrements étaient identiques pour cet identifiant (NJF1=N2 et N1=NJF2) de la même façon que si nous avions eu N1=N2 et NJF1=NJF2. La même méthode a été utilisée pour la prise en compte des deux prénoms.

Bases de données

Cook et coll (4) ont montré l'impact de la taille du fichier et de la probabilité de la corrélation sur le résultat de la corrélation. La corrélation serait plus facile lorsque l'on croise des fichiers de petite taille et correspondant à des populations proches. La corrélation serait donc également plus aisée entre deux fichiers d'un même réseau comprenant des patients en nombre limité et beaucoup de patients communs. Au contraire, dans le cadre d'échanges de données entre établissements n'étant pas liés par des conventions de coopération, la probabilité qu'un patient soit hospitalisé dans les deux structures est faible, ce qui diminuerait la qualité de la corrélation.

Il était donc nécessaire de prendre en compte différentes situations selon la taille des fichiers et la probabilité de corrélation.

- situation 1 : corrélation des données de deux années consécutives (100 000 enregistrements chacun) d'un centre hospitalier universitaire (CHU de Dijon)
- situation 2 : corrélation des données de deux années consécutives (50 000 enregistrements chacun) d'un hôpital de taille moyenne (établissement de l'APHP)
- situation 3 : corrélation de données provenant d'un établissement hospitalier périphérique de petite taille (200 enregistrements) et de données provenant d'un centre hospitalier universitaire (2500 enregistrements). Cette dernière situation permet de simuler une corrélation de fichiers dans le cadre d'un réseau de soins (dans notre cas le réseau périnatal de Bourgogne). Dans cette application, l'objectif est de corréler les données maternelles quel que soit leur lieu d'hospitalisation.

Dans les deux premières situations (corrélation de données provenant d'un même établissement), nous avons choisi de corréler les données de deux années consécutives afin d'éviter que la probabilité de retrouver un patient dans les deux fichiers soit nulle.

Pour respecter les législations françaises et européennes concernant le croisement des fichiers nominatifs, une technique d'anonymat (utilisant le Standard Hash Algorithm) a été appliquée aux bases testées, dès leur extraction dans les établissements sources. Sur le plan méthodologique, nous avons déjà démontré qu'il est équivalent, pour l'évaluation de la qualité de la corrélation, de prendre en compte des bases avec des données en clair ou rendues anonymes par hachage (7).

Méthode

Statistique

L'objectif de la corrélation est de confronter des fichiers doublement hachés provenant de sources différentes, pour associer les observations qui se rapportent à un même individu. Deux types d'erreurs (8) peuvent survenir dans le processus de corrélation. Le premier correspond à la corrélation de deux observations concernant deux individus différents et constitue une erreur « d'homonymie » (collision). Ce type d'erreur survient par exemple si l'on associe à tort des informations concernant deux personnes dénommées respectivement Dupond et Dupont, du fait d'une erreur dans la saisie de leurs identités. Le deuxième type d'erreur correspond à l'absence de corrélation de deux observations d'un même individu et constitue l'erreur « de synonymie » (doublon). Ce type d'erreur survient par exemple en cas d'utilisation successive du nom de jeune fille et du nom marital pour la même femme.

Nous avons adapté (6, 9, 10) la méthode de corrélation "AUTOMATCH" proposée par JARO (11) et très utilisée aux USA (12). Cette méthode permet de prendre en compte simultanément plusieurs traits d'identification. Bien sûr, aucun de ces traits d'identification n'identifie un individu de manière pathognomonique, et l'on est ramené au problème connu de la valeur informationnelle d'un signe. Chaque trait d'identification est pondéré en fonction de la

quantité d'information qu'il apporte. Par exemple, on attribue une valeur plus importante à l'information fournie par la date de naissance qu'à celle fournie par le sexe. Pour déterminer si deux observations doivent être corrélées, on applique un modèle d'analyse statistique qui tient compte des coefficients de pondération de chaque trait d'identification utilisé. Le poids pour chaque trait d'identification i est : $W_i = \log(m_i/u_i)$, ce qui correspond au logarithme du rapport de vraisemblance positif (m_i/u_i) utilisé classiquement par les statisticiens, où les paramètres m_i et $(1-u_i)$ représentent respectivement la sensibilité et la spécificité du trait d'identification considéré. Ceci signifie que le taux de doublons = $1-m_i$, et que le taux de collisions = u_i . La décision à prendre pour classer une paire d'enregistrements dépend de l'ensemble des traits d'identification. Ainsi, on attribue à chaque paire d'enregistrements (Figure 1) un poids global appelé poids composé W_k égal à la somme des poids correspondant aux différents traits d'identification. Pour chaque trait d'identification, ce poids est positif en cas de concordance des deux enregistrements et négatif en cas de discordance.

	Nom	Prénom	Date de naissance	Sexe	
	DUPONT	François	29-01-40	M	
	DUPOND	François	29-03-40	M	
Poids	+ 9	+ 4	- 3	+ 2	= 12

Figure 1. Exemple du calcul du poids composé lors de la corrélation de deux enregistrements.

A partir de la méthode ci-dessus, on peut donc estimer le rapport de vraisemblance ou poids W_i associé à chaque trait, sa sensibilité et spécificité ainsi que les taux de doublons et de collisions correspondants. On pourra ainsi sélectionner les traits de forte valeur discriminante (poids très élevés).

RESULTATS

Les résultats sur les données de Dijon (situation 1 - Tableau I) montrent que le classement des critères par valeur discriminante estimée par le rapport de vraisemblance positif est en faveur de la date de naissance ($r = 10,16$) en deuxième position vient le nom puis le prénom.

Tableau I. Estimation des valeurs discriminantes des traits d'identification (données du CHU de Dijon, avec traitement phonétique adapté à la langue française)

Trait	(m) : sensibilité 1 – taux doublons (%)	(u) : 1 – spécificité taux de collisions (%)	Logarithme du rapport de vraisemblance $w=(\ln(m/u))$
Nom usuel	99,7	2.26 E-02	8,39
Prénom	97,8	4.47 E-01	5,39
Date de naissance	99,7	3.87 E-03	10,16
Sexe	100	50	0,69

La sensibilité du nom est aussi élevée que celle de la date de naissance (sensibilité 99,7%, taux de doublons = 3×10^{-3}) mais la spécificité est moins bonne. En effet, le taux de collision (1- la spécificité) est de $2,26 \times 10^{-2}$ % pour le nom alors qu'il n'est que de $3,87 \times 10^{-3}$ % pour la date de naissance. Le sexe a un pouvoir discriminant qui est quasiment nul. Ceci s'explique essentiellement par la spécificité qui est de l'ordre de 50% : ainsi les enregistrements de deux personnes différentes ont une probabilité de 50% d'avoir le même sexe. Le rapport de vraisemblance, qui tient compte à la fois de la sensibilité et la spécificité, est proche de 0.

Par conséquent, **dans tous les traitements ultérieurs**, seuls la date de naissance, le nom et le prénom ont été considérés.

L'application du ou inclusif (Tableau II) entre le nom usuel et le nom de jeune fille ne semble rien apporter : la sensibilité du nom usuel étant déjà très élevée (99,7 %), le gain apporté par la prise en compte du nom de jeune fille n'apparaît pas. Comme l'on pouvait s'y attendre, la spécificité diminue dans le cas du nom composé. L'apport de l'information sur le nom de jeune fille aurait peut être été plus important si l'on avait considéré deux établissements différents (ayant des règles d'identification différentes) pour deux périodes moins rapprochées (augmentant la probabilité de changement de nom).

Tableau II. Estimation des valeurs discriminantes des traits d'identification en fonction des différents traitements phonétiques possibles (données du CHU de Dijon)

	(m) : sensibilité 1 – taux doublons (%)	(u) : 1 – spécificité taux de collisions (%)	Logarithme du rapport de vraisemblance (ln(m/u))
<u>Avec traitement phonétique adapté à la langue française</u>			
Nom usuel	99,8	2,27 E-02	8,39
Premier prénom	97,8	4,47 E-01	5,39
Date de Naissance	99,9	3,87 E-03	10,16
Nom composé *			
Nom composé *	99,8	3,80 E-02	7,87
Premier prénom	97,8	4,47 E-01	5,39
Date de Naissance	99,9	3,87 E-03	10,16
<u>Avec traitement par SOUNDEX</u>			
Nom composé*	99,9	2,16 E-01	6,14
Premier prénom	99,3	7,06 E-01	4,95
Date de Naissance	99,9	3,87 E-03	10,16
<u>Sans traitement phonétique</u>			
Nom composé*	99,9	3,14 E-02	8,07
Premier prénom	97,8	4,15 E-01	5,46
Date de Naissance	99,9	3,88 E-03	10,16

*Nom composé : nom usuel et nom de jeune fille combinés par un ou inclusif

Si l'on compare les résultats obtenus avec le traitement phonétique adapté à la langue française et ceux obtenus avec le SOUNDEX, on s'aperçoit que la sensibilité du prénom augmente avec le SOUNDEX (ce qui signifie une réduction du taux de doublons). Par contre, le taux de collision augmente de manière plus marquée pour le nom que pour le prénom (ce qui se traduit par une diminution de la spécificité). Au total, les valeurs discriminantes du nom et du prénom sont donc moins élevées avec le SOUNDEX qu'avec le traitement francisé.

La comparaison des résultats avec le traitement phonétique adapté à la langue française versus sans traitement phonétique montre peu de différences, si ce n'est une légère diminution de la spécificité avec le traitement phonétique sans amélioration de la sensibilité. Ce résultat doit être interprété avec précautions dans la mesure où l'apport du traitement phonétique ne peut être évalué de façon fiable que lors du croisement de bases de données provenant d'établissements différents (qualité et règle de saisies différentes suivant les établissements). Or, cette étude n'a pu être réalisée que sur les

données du CHU de Dijon. En effet, les autres situations impliquaient d'autres établissements que le nôtre et les données étaient rendues anonymes dès leur extraction.

Quelle que soit la situation considérée, en utilisant le traitement phonétique adaptée à langue française (Tableau III), la classification des différents traits d'identification reposant sur leur valeur discriminante (rapport de vraisemblance) n'est pas modifiée. La date de naissance est toujours le trait le plus discriminant, suivi par le nom (combinaison par un ou inclusif du nom usuel et du nom de jeune fille), puis par le premier prénom.

Tableau III. Estimation des valeurs discriminantes des traits d'identification (résultats dans les 3 situations après traitement phonétique adapté à la langue française)

Traits d'identification	(m) : spécificité 1 – taux de doublons (%)	(u) : 1 – spécificité taux de collision(%)	Proportion d'enregistrements à chaîner(%)	Logarithme du rapport de vraisemblance $w=(\ln(m/u))$
1ère situation : CHU de Dijon				
Nom composé*	99.80	3.80 E-02	2.17 E-02	7.87
Premier prénom	97.80	4,47 E-01	2.17 E-02	5.39
Date de naissance	99.90	3.87 E-03	2.17 E-02	10.16
2ème situation : hôpital de taille moyenne				
Noms composé*	95.70	9.33	3.40 E-04	9.24
Premier prénom	98.90	2.30 E-01	3.40 E-04	6.06
Date de naissance	85.40	3.50 E-03	3.40 E-04	10.10
Nom composé*	93.00	9.31 E-03	3.75 E-04	9.21
Prénom composé**	99.00	2.30 E-01	3.75 E-04	6.06
Date de naissance	79.70	3.49 E-03	3.75 E-04	10.04
3ème situation : réseau périnatal de Bourgogne				
Nom de jeune fille	98.70	1.16 E-01	2.66 E-02	6.75
Premier prénom	79.10	1.24	2.66 E-02	4.16
Date de naissance	100.00	1.46 E-02	2.66 E-02	8.83

* nom usuel et nom de jeune fille combinés par un ou inclusif

** premier et deuxième prénoms combinés par un ou inclusif

Lorsqu'on considère une corrélation de deux fichiers d'un hôpital de taille moyenne (situation 2) par rapport à une corrélation de deux fichiers issus d'un CHU, on s'aperçoit que la valeur discriminante du nom et du prénom augmente. Au contraire, la valeur discriminante de la date de naissance diminue légèrement. Ce phénomène peut être partiellement expliqué par la différence du pourcentage d'enregistrements à chaîner dans les deux établissements. Il peut être également relié à la différence de proportion de sujets d'origine étrangère dans les deux hôpitaux, l'hôpital de taille moyenne étant localisé à Paris. Une autre raison peut être également avancée : la qualité de la saisie des données n'est peut être pas non plus étrangère à ces résultats.

L'ajout du 2^{ème} prénom dans le traitement, n'a été possible que sur les données parisiennes où les deux prénoms étaient parfois renseignés. On ne constate pas d'amélioration de la valeur discriminante lors de l'ajout du 2^{ème} prénom (taux de doublons, de collision ou sensibilité, spécificité quasiment identiques) par rapport au seul traitement du 1^{er} prénom. D'autre part, le modèle paraît perturbé, ce qui se traduit par une augmentation de la proportion des paires à chaîner et une diminution de la sensibilité des autres critères (nom, date de naissance). Sur le plan statistique, ce modèle est donc moins bon, ce qui signifierait que l'ajout du 2^{ème} prénom diminue la valeur discriminante de l'ensemble des traits utilisés. L'explication de ce résultat vient peut être du fait que le 2^{ème} prénom est rarement renseigné (au maximum une fois sur deux). Par contre, quand il est renseigné, l'information apportée n'est probablement pas discriminante dans la mesure où cela peut conduire à rapprocher à tort plus de paires d'enregistrements. Du fait de l'application du ou-inclusif entre les deux prénoms, on pourrait alors être amené à considérer comme identiques deux individus qui sont peut être différents.

Lorsqu'on considère la situation 3 (réseau périnatal de Bourgogne), la valeur discriminante des trois identifiants (nom, prénom, date de naissance) est légèrement réduite par rapport à ce qui est observé dans la situation 1. La réduction observée pour la date de naissance n'est pas très étonnante car l'âge des femmes qui accouchent varie dans une fourchette beaucoup moins large que pour l'ensemble des hospitalisés au CHU (le taux de collision augmente). De la même façon, le fait de se limiter aux prénoms des femmes va limiter l'éventail des prénoms possible. On constate donc un effet sur le taux de collision, mais aussi une augmentation du taux de doublons qui est peut être liée à la moindre qualité des données liées à la diversité des sources. En effet, les données de la périnatalité ne proviennent pas que du CHU mais de l'ensemble des établissements de la région et font intervenir les différences de qualité et de règles du recueil de ces établissements.

DISCUSSION

Les conclusions de ce travail montrent l'intérêt de la prise en compte de la date de naissance, du nom et du prénom pour le regroupement des données d'un même patient.

Cette étude a également montré que le fait de rajouter une variable supplémentaire peu discriminante, telle que le sexe, n'apporte rien à la qualité de ce regroupement. En effet, ce trait est très peu spécifique, ce qui signifie que le taux de collisions est très important (de l'ordre de 50%) : deux enregistrements provenant de deux patients différents ont une probabilité de 50% d'avoir le même sexe. On aurait pu se demander si la prise en compte du sexe permettait de différencier les enregistrements correspondant à deux patients de même date de naissance, nom et prénom, lorsqu'un doute subsiste à la lecture du prénom (Dominique par exemple). L'application du modèle montre que l'information apportée par le sexe est négligeable par rapport à celle sur la date de naissance, le nom et le prénom. En effet, lorsque deux enregistrements ont la même date de naissance, le même nom et le même prénom, la proportion observée sur les données dijonnaises pour que ces deux enregistrements n'aient pas le même sexe est très faible (de l'ordre de $2 \cdot 10^{-10}$).

D'autre part ces résultats montrent que l'utilisation de l'information sur le 2^{ème} prénom n'améliore pas la qualité du chaînage. On peut noter également, que la prise en compte simultanée (sous la forme du nom composé) des deux informations (nom usuel et nom de jeune fille) ne permet pas d'améliorer la corrélation obtenue avec une seule information. Toutefois, l'intérêt de ces critères serait peut-être plus évident sur d'autres bases de données, pour lesquelles un changement de nom ou de règles d'identification entre les deux fichiers corrélés seraient plus fréquents que lors du croisement de deux années consécutives d'un même établissement.

Ce travail a montré que la comparaison du traitement phonétique francisé et du traitement par SOUNDEX est légèrement en faveur du traitement phonétique francisé. Nous avons montré dans d'autres travaux (soumis pour publication) que l'apport d'un traitement phonétique pour permettre le rapprochement de deux identités mal orthographiées est plus important dans le cas d'un chaînage direct (sans modèle probabiliste). En effet, si les deux noms sont orthographiés différemment, le chaînage ne sera alors pas possible. Dans le cas d'un modèle probabiliste, l'égalité des autres

champs (dates de naissance et prénoms) permettra d'envisager le chaînage même si tous les enregistrements ne sont pas rigoureusement identiques.

Les recommandations qui peuvent être apportées à l'issue de ce travail sont surtout d'ordre pragmatique. Pour permettre un regroupement des données d'un même patient dans les meilleures conditions, il serait préférable de se limiter au recueil des trois traits précédemment cités (date de naissance, nom et prénom). Pour ce qui concerne le nom, le choix entre le nom usuel ou le nom de jeune fille repose sur la disponibilité de l'information, notamment dans les pièces administratives. L'intérêt du nom de jeune fille ou plutôt du nom de naissance, est d'être stable au cours du temps. Il serait, par ailleurs, préférable de limiter le recueil au 1^{er} prénom figurant sur la pièce administrative plutôt que de vouloir systématiser le recueil des prénoms multiples.

Il serait également souhaitable de se focaliser sur la qualité du recueil et de vérifier chaque fois que cela est possible, l'exactitude de la date de naissance, du nom et du prénom, à partir par exemple d'une pièce administrative d'identité (carte d'identité ou carte vitale ?). Il pourrait être aussi recommandé d'appliquer une procédure de validation, notamment sur le nom et le prénom pour permettre de réduire les erreurs de saisie.

REMERCIEMENTS

Cette étude a été initiée par une Groupe de travail du service en charge de la Modernisation du Système d'Information Hospitalier (GMSIH.) et a été confié au Département d'Information Médicale (DIM) du CHU de Dijon.

Nous tenons à remercier Mr. Marie, Directeur Général du CHU de Dijon ainsi que le Directeur du système d'information, Mr. Bitouzé pour avoir autorisé l'exploitation des données. Nous voulons également remercier Mr. Pontou, qui a la charge de l'unité administrative de gestion des patients au sein du service de gestion du système informatique du CHU de Paris (AP-HP). Nous tenons aussi à remercier particulièrement les membres du réseau périnatal de Bourgogne pour leur collaboration précieuse, de même que Melle Harmenil et Mr. Pequignot pour leur importante contribution à l'exploitation des données.

REFERENCES

1. Newcombe H. Handbook of record linkage: methods for health and statistical studies, administration and business. *Oxford University Press ed. New York*; 1998.
2. Brenner H, Schmidtman I. Determinants of homonym and synonym rates of record linkage in disease registration. *Meth Inform Med* 1996;35:19-24.
3. Brenner H, Schmidtman I. Effects of record linkage errors on disease registration. *Meth Inform Med* 1998;35:19-24.
4. Cook L, Olson L, Dean J. Probabilistic record linkage: relationships between file sizes, identifiers, and match weight. *Meth Inform Med* 2001;40:196-203.
5. Thirion X, Sambuc R, San-Marco J. Epidemiology and anonymity: a new method. *Rev Epidemiol Sante Pub* 1988;36:36-42.
6. Quantin C, Bouzelat H, Allaert F, Benhamiche A, Faivre J, Dusserre L. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Meth Inform Med* 1998;37:271-7.
7. Quantin C, Allaert F, Pattisina R, Biquet C, Cornet B, Gouyon J, et al. Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat : application au suivi de informations médicales. *VIIèmes Journées de Méthodologie Statistique*. Paris, France, 4 et 5 décembre 2000.
8. Brenner H, Schmidtman I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. *Stat Med* 1997;16:2633-43.
9. Quantin C, Bouzelat H, Allaert F, Benhamiche A, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow-up: quality assessment of an homonymous record linkage procedure. *Int J Med Informatics* 1998;49:117-22.

10. Quantin C, Allaert F, Bouzelat H, Rodrigues J, Trombert-Paviot B, Brunet-Lecomte P, et al. La sécurité des réseaux d'informations médicales : application aux études épidémiologiques. *Rev Epidemiol Sante Pub* 2000;48:89-99.
11. Jaro M. Probabilistic linkage of large public health data files. *Stat Med* 1995;14:491-8.
12. Sugarman J, Hollyday M, Ross A, Castorina J, Hui Y. Improving american indian cancer data in the Whashington state cancer registry using linkages with the Indian health service and tribal records. *Am Cancer Soc* 1996;78:1564-8.
13. Pequignot S. Anonymisation et chainage de données hospitalières [*Mémoire de DESS "Base de données et Intelligence artificielle"*]: Université de Bourgogne; 2001.