ANNEXE 2
Guide d'utilisation du programme SALI : un logiciel de couplage de données

Table des matières

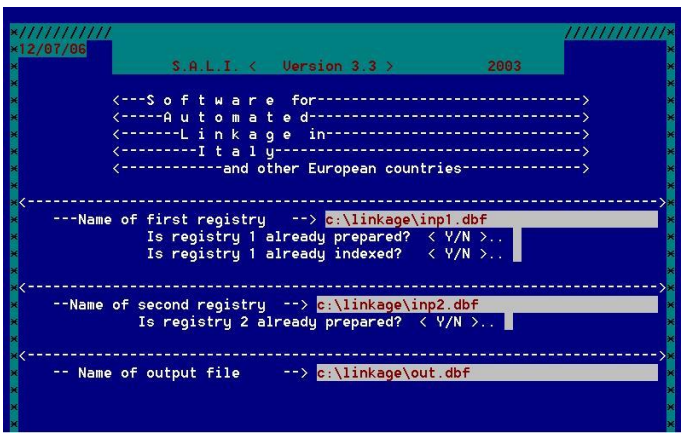
ANNEXE 2
Guide d'utilisation du programme SALI : un logiciel de couplage de données

Instructions pour le couplage à l'aide de SALI

Les fichiers de données à assembler doivent posséder le format .DBF (DBIII Plus ou DBIV) et contenir un numéro d'identification unique.
Les dates doivent être converties en champs séparés pour les informations sur le jour, le mois et l'année (pour le format "chaîne de caractères" se reporter à la capture d'écran suivante).
Le programme gère également les doublons, ce qui implique que les fichiers à unir ne doivent pas nécessairement contenir un seul enregistrement pour un même nom, prénom et date de naissance.
En revanche, cette situation peut rallonger le temps prévu pour l'opération de couplage.
Il est nécessaire que chaque enregistrement soit identifié à travers un unique code (comme par exemple un numéro progressif), de manière à ce que le second enregistrement identifié à l'aide du même code (suivant le premier) ne soit pas considéré dans la procédure de couplage. Pour accélérer la procédure, le fichier contenant le plus grand nombre d'enregistrements devrait correspondre au ficher numéro 1.
Un exemple de la structure du fichier est présentée ici :

Fichier 1	
ID1	numéro d'identification unique de l'enregistrement dans le fichier 1
NOM1	nom
PRENOM1	prénom
JNAIS1	jour de naissance
MNAIS1	mois de naissance
ANAIS1	année de naissance
VAR1 …	variables additionnelles fichier 1
Fichier 2	
ID2	numéro d'identification unique de l'enregistrement dans le fichier 2
NOM2	nom
PRENOM2	prénom
JNAIS2	jour de naissance
MNAIS2	mois de naissance
ANAIS2	année de naissance
VAR2 …	variables additionnelles fichier 2

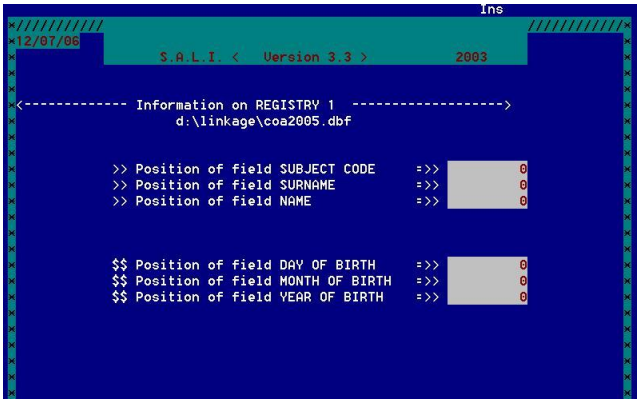
Les fichiers ainsi préparés doivent être positionnés dans un dossier (par exemple c:\couplage) ; il est préférable (mais pas nécessaire) de placer les fichiers et le programme dans le même dossier.
Une petite fenêtre apparaît alors où il est demandé d'indiquer le chemin correspondant aux fichiers 1 et 2 (reconnaissables par l'extension .dbf) ainsi que le répertoire où devra être sauvegardé le fichier de sortie (ou d'output, par exemple out.dbf).



Au moment de l'insertion du 1er fichier, le programme demande :

- si le fichier n'a pas été préparé (noms et prénoms en lettres majuscules sans utiliser ni espace ni signes de ponctuation ; par exemple, "Da Vinci" et "Maria-Josè" doivent apparaitre écrits DAVINCI et MARIAJOSE) : répondre "n" ;
- si le fichier n'a pas été trié sur toutes les clés primaires : répondre "n".

De même, lors du chargement du 2nd fichier, le programme demande si le fichier a été préparé : répondre "n".
Si tout a été correctement effectué (en l'absence de message de fichier non existant), après avoir appuyé sur la touche ENTREE, le programme requiert l'insertion de la position dans l'ordre séquentiel des champs qui devront être utilisés pour le couplage.



Une fois cette opération terminée, une demande d'autorisation à poursuivre les procédures apparaît. Le programme montre alors sept niveaux différents de couplage qui varient selon la modalité d'utilisation des variables, comme cela est montré dans le tableau 1 (*SALI* version 3.3).

Tableau 1

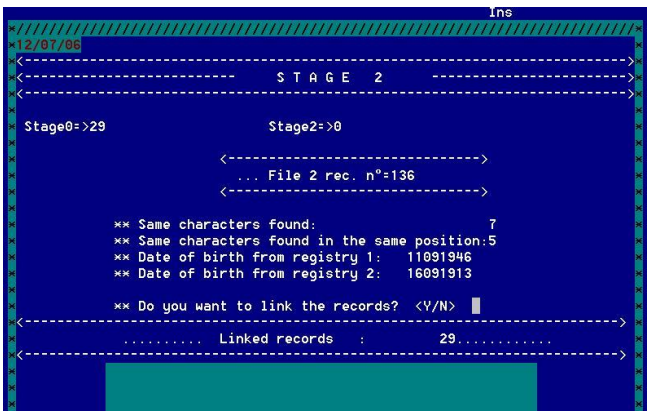
Niveau[a]	Nom	Prénom	Date de naissance	Intervention manuelle[b]
0	Identique	un des deux prénoms est contenu dans l'autre	Identique	Non
1[c]	Identique	identique au niveau 0 ou bien les 7 premiers caractères ou bien un prénom sans la première lettre est contenue dans l'autre	même année	non
2	Identique	identique au niveau 1	au moins 7 caractères en commun et au moins 5 caractères dans la même position	oui
3	pour chaque nom, un maximum de 20 caractères ne sont pas trouvés dans l'autre et une chaîne d'au moins 2 caractères en commun	identique au niveau 1	identique (mois et jour inversés)	oui
4	un des deux noms contenus dans l'autre ou bien les mêmes 7 premiers caractères pour chaque prénom, un maximum de 20 caractères non retrouvés dans l'autre et une chaîne d'au moins 2 caractères en commun	identique au niveau 3	oui	
5	un des deux noms contenus dans l'autre ou bien les mêmes 7 premiers caractères	identique au niveau 4	au moins 7 caractères en commun et au moins 6 dans la même position	oui
6	pour chaque nom, un maximum de 3 caractères non retrouvés dans l'autre et une chaîne d'au moins 4 caractères en commun	un des 2 prénoms identique à la première moitié de l'autre	identique au niveau 5	oui

[a] les couplages effectués lors d'un niveau précédent s'annulent à partir du niveau suivant
[b] en partant du niveau 2, tous les couplages possibles sont proposés l'un après l'autre et l'opérateur doit décider si l'accepter ou non
[c] à cause d'une moins bonne spécificité, dans la version SALI 3.3, le niveau 1 est proposé après le niveau 5

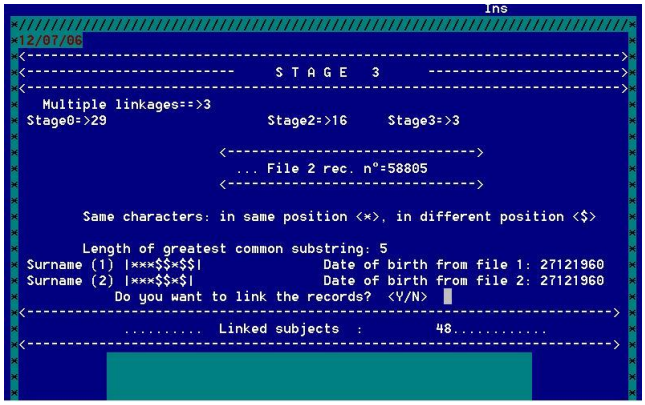
Autres fonctionnalités du logiciel

Il est possible de sauter un niveau proposé.
Les noms et prénoms peuvent être supprimés, recodés (cryptés), ou laissés inchangés dans le fichier de sortie.
Le niveau 6 ne peut être utilisé que dans des circonstances particulières lorsqu'une sensibilité élevée est nécessaire (procédure lourde).
Pour des raisons de confidentialité, bien qu'il s'agisse de fichiers contenant des identifiants de personnes, le programme n'affiche jamais de noms ou prénoms au cours de la procédure de couplage.
Aux niveaux permettant le choix manuel, le programme propose des cas similaires (voir le tableau 1) et assiste l'opérateur dans le choix de couplages basés sur deux chaînes cryptées de noms de même longueur.

A l'étape 2 (pour les noms et prénoms qui sont presque les mêmes, voir le tableau 1), la fenêtre suivante s'affiche :



A l'étape 3 (pour les mêmes dates de naissance, voir le tableau 1), la fenêtre suivante s'affiche :



où:
le symbole * signifie "mêmes caractères situés dans la même position ;"
le symbole $ signifie "mêmes caractères trouvés dans l'autre chaîne de caractères, mais dans une position différente ;"
le symbole - signifie "caractère différent."
Dans l'exemple décrit ci-dessus, les deux enregistrements ont le même prénom (non montré, voir le tableau 1), la même date de naissance, et une chaîne commune de 5 caractères dans le nom de famille dans les positions indiquées (un exemple classique d'un double caractère comme erreur de transcription dans le nom de famille).

Il convient de noter que :

- seule l'astérisque signifie qu'il existe une correspondance entre les chaînes, tandis que les autres symboles, bien qu'ils produisent un effet optique similaire, peuvent être associés tout en comportant des différences considérables ;
- si le choix est "n" (NO) à la demande de couplage, les deux cas seront exclus; Toutefois, après avoir choisi "y" (OUI), il sera possible de rejeter le couplage à une phase de contrôle successive lors de la comparaison des champs de données supplémentaires qui se trouvent dans les deux archives (ex. : la ville de naissance, la date de décès) ;
- Les différentes conditions de fonctionnement (la taille des fichiers à coupler, la probabilité de petites différences correspondant en réalité à différents patients) permettra de déterminer la pertinence de plus stricts (fonctionnement plus rapide, plus grande spécificité) ou plus flexibles (fonctionnement plus lent, plus grande sensibilité) critères de couplage.

A la fin de la procédure, le programme demande si les chaînes avec les noms et prénoms doivent être supprimés : cette option est obligatoire en cas de couplage à effectuer "en aveugle" pour des raisons de confidentialité. Dans ce cas, afin de permettre un contrôle ultérieur, seul les nom et prénom des chaînes cryptées à l'aide *, $ et - seront conservés ; autrement, si les fichiers peuvent être librement utilisés par les opérateurs, le fichier de sortie, en plus de tous les champs de données des enregistrements, contiendra également les noms et prénoms des deux fichiers dans une forme non cryptée.
Après la fin de la procédure, un test final des enregistrements couplés (afin d'éliminer les "faux positifs") peut être effectué en utilisant n'importe quel type de logiciel (DB3/4, Excel, Access, etc.) tout en gardant en mémoire que chaque enregistrement contenant les données des deux fichiers contient également un champ de données informant sur le "niveau" de couplage réalisé. Si des critères de couplage plus généreux devaient être choisis (à titre d'exemple extrême, si "y" est choisi en réponse à chaque demande), il sera possible de sélectionner tous les enregistrements avec un niveau de couplage qui est, par exemple, supérieur à 1, afin de re-vérifier et de les accepter/rejeter à ce moment-là.

Le logiciel est disponible gratuitement et uniquement à des fins de recherche épidémiologique, sur demande écrite à :

Dr. Luigino Dal Maso, Unité d'Epidémiologie et de Biostatistique (e-mail : epidemiology@cro.it)
Centro Riferimento Oncologico, via Pedemontana occ. 12, 33081 Aviano (Pn)

Lors de l'utilisation de ce logiciel, il est prié d'indiquer la référence bibliographique ci-dessous.

Bibliographie

Dal Maso L, Braga C, Franceschi S. Methodology used for "Software for automated linkage in Italy" (SALI). Journal of Biomedical Informatics 2001; 34: 387-95.