APPENDIX 2

Contents

APPENDIX 2
User's Guide to SALI : Software for Record Linkage

Instructions for linkage using *SALI*

The files must be in .DBF (DBIII Plus or DBIV) format and must contain a unique identification code.

The dates must be converted into separate fields for day, month, and year (for character format, see the following layout).

Since the software can also work with duplicates, the files to be linked do not need to contain the patient's surname-name-DOB only once. This, however, may make the manual linkage phase longer.

It is nevertheless necessary for each record to have a unique identifier (e.g., a sequential number), otherwise records with the same identifier (subsequent to the first one) will not be taken into consideration.

To speed up the procedure, the file with the greatest number of records should be file *1*.

Here is an example of the structure:

File1

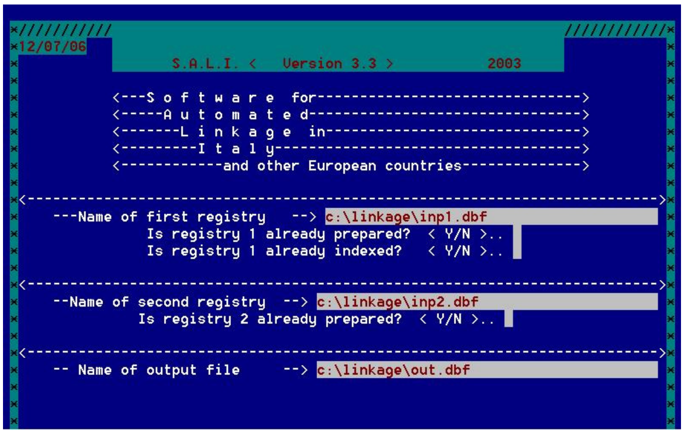
ID1	unique identifier of the record in file 1
COGN1	surname
NOME1	name
GGNAS1	day of birth
MMNAS1	month of birth
AAAANAS1	year of birth
VAR1 …	additional *file*1 variables

File2

ID2	unique identifier of the record in file 2
COGN2	surname
NOME2	name
GGNAS2	day of birth
MMNAS2	month of birth
AAAANAS2	year of birth
VAR2 …	additional file 2 variables

When ready, the files must be placed in a folder (e.g., c:\linkage); it is preferable (but not necessary) to place files and software in the same folder.

A small window will appear, where you will need to specify the address of file 1 and file 2 (the files are to be designated by the extension .dbf) and where you want to place the output file (e.g., out.dbf).



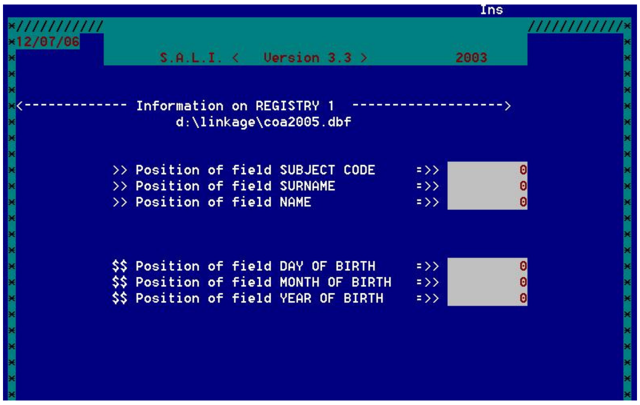
```
12/07/06
        S.A.L.I. <   Version 3.3 >            2003

      <---S o f t w a r e  for------------------------------>
      <-----A u t o m a t e d------------------------------->
      <-------L i n k a g e  in----------------------------->
      <---------I t a l y----------------------------------->
      <------------and other European countries------------->

<------------------------------------------------------------------------>
   ---Name of first registry   --> c:\linkage\inp1.dbf
          Is registry 1 already prepared?  < Y/N >..
          Is registry 1 already indexed?   < Y/N >..

<------------------------------------------------------------------------>
   --Name of second registry  --> c:\linkage\inp2.dbf
         Is registry 2 already prepared?  < Y/N >..

<------------------------------------------------------------------------>
   -- Name of output file     --> c:\linkage\out.dbf
```

When you insert file 1, the program asks:

- whether the file has been prepared (names and surnames in capital letters with no spaces or punctuation marks; e.g., “Da Vinci” and “Maria-Josè” become DAVINCI and MARIAJOSE): answer “**n**”;
- whether the file has not been indexed (sorted) on all the key variables: answer “**n**”.

When you insert file 2, the program asks whether the file has been prepared: answer “**n**”.

If there are no problems (e.g., non-existent files), after pressing ENTER the program will prompt you to indicate the position (sequential order) of the data fields that will be used for linkage in the files.



```
                                                                     Ins
*///////////                                                        ///////////*
*12/07/06
*
*          S.A.L.I. <   Version 3.3 >            2003

*<------------ Information on REGISTRY 1  ------------------>
                 d:\linkage\coa2005.dbf

               >> Position of field SUBJECT CODE      =>>          0
               >> Position of field SURNAME           =>>          0
               >> Position of field NAME              =>>          0


               $$ Position of field DAY OF BIRTH      =>>          0
               $$ Position of field MONTH OF BIRTH    =>>          0
               $$ Position of field YEAR OF BIRTH     =>>          0
```

Once this is done, the program will request permission to execute the procedures.

Next, you will be shown seven different levels of linkage with different treatment of variables, as shown in Table 1 (*SALI* version 3.3).

Table 1

Level[a]	Surname	Name	Date of Birth	Manual Intervention[b]
0	Same	one of the two names contained in the other	same	no
1[c]	Same	same as at level 0, <u>or</u> same first 7 characters, <u>or</u> a name without the first letter contained in the other	same year	no
2	Same	same as at level 1	at least 7 characters in common and at least 5 characters in the same position	yes
3	for each surname, a maximum of 20 characters not found in the other and a string of at least 2 characters in common	same as at level 1	same (month and day may be inverted)	yes
4	one of the two surnames contained in the other or same first 7 characters	for each name, a maximum of 20 characters not found in the other and a string of at least 2 characters in common	same as at level 3	yes
5	one of the two surnames contained in the other or same first 7 characters	same as at level 4	at least 7 characters in common and at least 6 in the same position	yes
6	for each surname, a maximum of 3 characters not found in the other and a string of at least 4 characters in common	one of the two names identical to the first half of the other	same as at level 5	yes

[a] *matches made in a preceding stage are disregarded starting from the next*

[b] *starting from level 2, all the possible linkages are proposed one by one and the operator must decide whether to accept them or not*

[c] *due to lesser specificity, in SALI 3.3, level 1 is proposed after level 5*

Other features of the software

You can skip the proposed level.

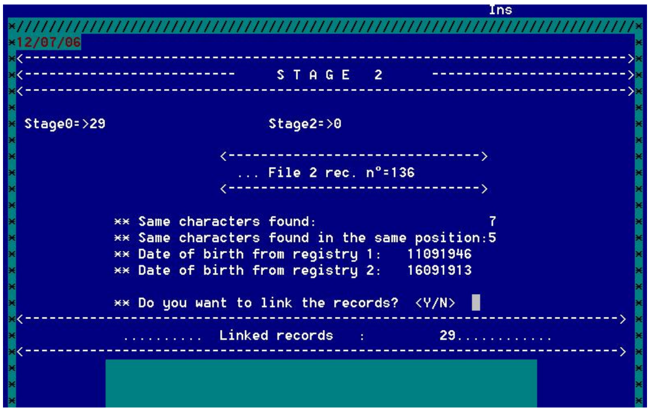
Names and surnames can be deleted, encrypted, or left unchanged in the output file.

Level 6 can be used only in special circumstances when high sensitivity is required (onerous procedure).

Due to privacy considerations, though using files containing personal names, the program never displays names or surnames during the linkage stage.

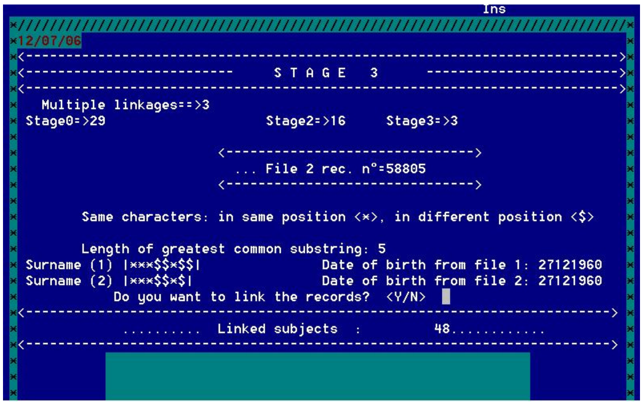
At the levels allowing for manual choice, the registrar will be presented with similar cases (see Table 1) and assisted in choosing whether the linkages should be accepted or not based on two encrypted strings of names with the same character length.

In Stage 2 (for names and surnames that are *almost* the same , see Table 1), the following window will be displayed:



```
                                                                 Ins
*///////////////////////////////////////////////////////////////////////////*
*12/07/06
*<------------------------------------------------------------------------->*
*<-------------------------     S T A G E   2     ------------------------->*
*<------------------------------------------------------------------------->*
*
*  Stage0=>29                     Stage2=>0
*
*                             <------------------------------->
*                               ... File 2 rec. n°=136
*                             <------------------------------->
*
*              ** Same characters found:                      7
*              ** Same characters found in the same position:5
*              ** Date of birth from registry 1:   11091946
*              ** Date of birth from registry 2:   16091913
*
*              ** Do you want to link the records?  <Y/N>
*<------------------------------------------------------------------------>*
*               .........  Linked records   :         29...........
*<------------------------------------------------------------------------>*
```

In Stage 3 (for same dates of birth, see Table 1), the following window will be displayed:



```
                                                                 Ins
*///////////////////////////////////////////////////////////////////////////*
*12/07/06
*<------------------------------------------------------------------------->*
*<-------------------------     S T A G E   3     ------------------------->*
*<------------------------------------------------------------------------->*
*   Multiple linkages==>3
*  Stage0=>29                     Stage2=>16     Stage3=>3
*
*                             <------------------------------->
*                               ... File 2 rec. n°=58805
*                             <------------------------------->
*
*          Same characters: in same position <*>, in different position <$>
*
*          Length of greatest common substring: 5
*  Surname (1) |***$$*$$|                  Date of birth from file 1: 27121960
*  Surname (2) |***$$*$|                   Date of birth from file 2: 27121960
*              Do you want to link the records?  <Y/N>
*<------------------------------------------------------------------------>*
*               .........  Linked subjects  :         48...........
*<------------------------------------------------------------------------>*
```

where:

the symbol * stands for “same characters found in the same position;”

the symbol $ stands for “same characters found in the other string, but in a different position;”

the symbol - stands for “different character.”

In the example described above, the two records have the same name (not shown, see Table 1), the same date of birth, and a common substring of 5 characters in the surname in the positions shown (a classical example of a double character transcription error in the surname).

It should be noted that:

- only the coincidence of the asterisks means that there is a correspondence between the strings, while the coincidence of the other symbols, albeit producing an optical similarity effect, may be associated with considerable differences;
- if you answer “n” (NO) to the linkage request, the two cases will be excluded; however, when answering “y” (YES) it will be possible to reject the linkage at a later checking stage when comparing additional data fields found in the two archives (e.g., town of birth, date of death);

- the various conditions of operation (the size of the files you want to link, the likelihood of small differences actually corresponding to different patients) will determine the appropriateness of stricter (faster operation, greater specificity) or more relaxed (slower operation, greater sensitivity) linkage criteria.

At the end of the procedure, the program asks whether the strings with the names and surnames should be deleted: this option is mandatory in case of linkage to be performed "blind" due to privacy considerations). In this case, in order to enable later checks, only the name and surname strings encrypted using *, $ and – will be retained; otherwise, if the files can be used freely by the operators, the output file, in addition to all the linked data fields of the common records, will also contain the names and surnames of the two files in unencrypted form.

After the procedure ends, a final test of the linked records (in order to eliminate "false positives") can be carried out using any type of software (DB3/4, Excel, Access, etc.) while bearing in mind that every record containing the data fields from both files will also contain a data field with the "level" of linkage performed. Should more generous criteria of inclusion be chosen (to give an extreme example, if "y" is pressed in response to every request), it will be possible to select records with a linkage level that is, for example, higher than 1, in order to re-check and accept/reject them at that point.

The software is available, free of charge and only for the purpose of epidemiological research, upon written request to:

Dr. Luigino Dal Maso, Unità di Epidemiologia e Biostatistica (e-mail epidemiology@cro.it)
Centro Riferimento Oncologico, via Pedemontana occ. 12, 33081 Aviano (Pn)

When using the software, please quote the bibliographical reference below.

Reference

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.