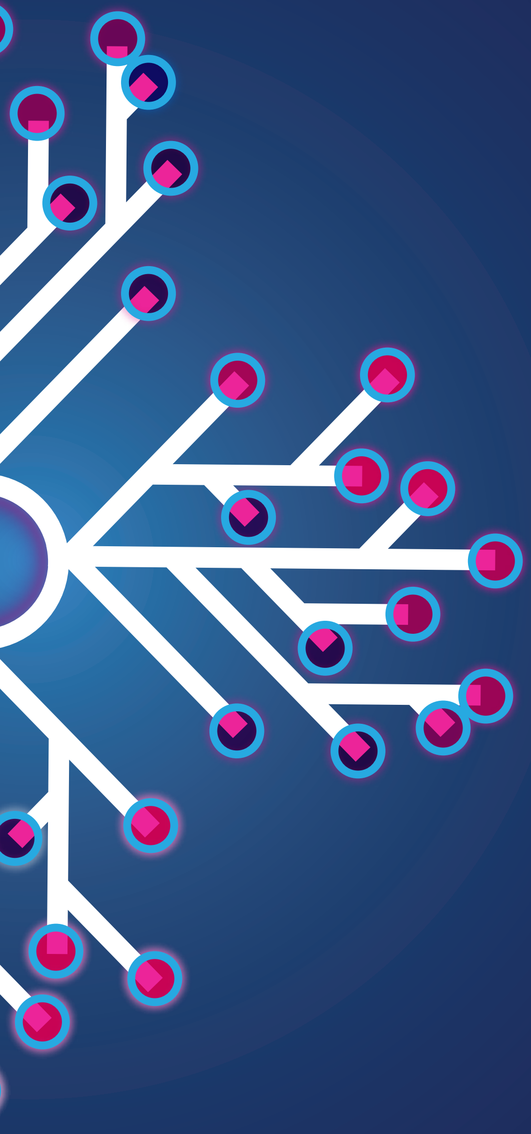
Consolidated guidance on tuberculosis data generation and use
Module 1

Tuberculosis surveillance

Web Annex C

Record-linkage exercises



World Health Organization

Consolidated guidance on tuberculosis data generation and use
Module 1

Tuberculosis surveillance

Web Annex C

Record-linkage exercises

World Health Organization

Consolidated guidance on tuberculosis data generation and use. Module 1. Tuberculosis surveillance. Web Annex C. Record-linkage exercises

ISBN 978-92-4-008091-1 (electronic version)

Suggested citation. Web Annex C. Record-linkage exercises. In: Consolidated guidance on tuberculosis data generation and use. Module 1: Tuberculosis surveillance. Geneva: World Health Organization; 2024. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Sales, rights and licensing. To purchase WHO publications, see https://www.who.int/publications/book-orders. To submit requests for commercial use and queries on rights and licensing, see https://www.who.int/copyright.

Third-party materials. If you wish to reuse material from this work that is attributed to a third party, such as tables, figures or images, it is your responsibility to determine whether permission is needed for that reuse and to obtain permission from the copyright holder. The risk of claims resulting from infringement of any third-party-owned component in the work rests solely with the user.

General disclaimers. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of WHO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific companies or of certain manufacturers' products does not imply that they are endorsed or recommended by WHO in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by WHO to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall WHO be liable for damages arising from its use.

This publication forms part of the document entitled Consolidated guidance on tuberculosis data generation and use. Module 1. Tuberculosis surveillance. It is being made publicly available for transparency purposes and information.

Contents

Chapter 1
Introduction

1.1 What is contained in this guidance?

This guide explains how to implement record linkage within the context of country programmes that are adopting digital case-based surveillance systems for tuberculosis (TB). It explains the importance of record linkage in finding missed TB cases and linking these to TB surveillance systems and, where possible, to TB care. There are three annexes:

- a description of how to use the Link Plus application to conduct probabilistic record linkage (Annex 1) and of how Link Plus can be used to deduplicate records (Annex 2); and
- further exercises to support readers intending to practise using Link Plus, and three comma separated variable (CSV) format files available from links provided (Annex 3).

1.2 Why external record linkage is important and where it is relevant?

Globally in 2022, it was estimated that about 3.1 million people (29% of incident TB patients)[1] were not detected or reported to the health systems. These "missing" patients were either undiagnosed or diagnosed but not reported. Poor linkage to treatment after diagnosis is one of the major contributors to "missing" patients.

With the establishment of digital case-based surveillance systems, people could be linked to effective treatment faster and more efficiently. For countries using digital TB surveillance systems, the World Health Organization (WHO) recommends implementing standard data validation algorithms to identify people who have been diagnosed but not linked to treatment. This could be done by matching individual TB notification and laboratory records through a process called record linkage, in which unique records are identified and information from both registers is then merged. Alternatively, TB notification data could be linked to civil registration and vital statistics records, to identify possible outcomes (e.g. mortality) among individuals who have been recorded as lost to follow-up. TB and HIV records may also be linked to ensure that people diagnosed with HIV and subsequently screened for TB according to country-specific treatment guidelines are also included in the TB surveillance system. At a population level, record linkage can help country programmes to clean up and deduplicate TB notification registers, and to effectively monitor the underlying burden of TB; it can also inform strategies aimed at making the TB surveillance system more robust. For operational research purposes, records may be linked across service delivery areas (e.g. clinic and inpatient records) to identify vulnerable subpopulations and their characteristics, for improved or more equitable programming of services.

Electronic patient management systems that are more advanced may have in-built functionalities to match laboratory and notification records, and to generate an automated list of nonmatched records. Where such functionality is not available, an alternative is to export TB notification and laboratory data in standard formats, then conduct external linkage.

If the separate systems use nationally defined unique personal identifiers (e.g. the social security number in the United States of America [USA]) in a systematic manner, then matching of records can be done directly. However, few countries use such identifiers, and even in those that do, direct linkage may be complicated when using names and birth dates, because of misspelling of names, or errors or missing values in birth dates. This guide explains **how to implement record linkage of two registers** to **identify the patients who were diagnosed but not included in the notification register.** The records are linked using the freely available probabilistic matching software Link Plus**.**

This guide is intended for use by national TB programme (NTP) data managers, epidemiologists and specialists in monitoring and evaluation (M&E) at the national and district level.

1.3 Concept of record linkage

Record linkage is used to establish whether two or more records relate to the same individual. It can be performed manually by visually comparing records; however, this is labour intensive, tedious and inefficient, particularly when the number of records increases.

[1] Global tuberculosis report 2023. Geneva: World Health Organization; 2023 (https://iris.who.int/handle/10665/373828).

Fig. WC.1.1 Example of matching records

Last Name	First Name	Region	SSN	DOB	Sex	Address
Omar	Mohammed	C619	123456789	23011966	1	P.O. Box 134
Omar	Mohammed	C619	123456789	23011966	1	P.O. Box 134

Fig. WC.1.2 Example of records containing variations

Last Name	First Name	Region	SSN	DOB	Sex	Address
Omar	Mohammed	C619	123456789	23011966	1	P.O. Box 134
Omar	Muhammed	C619	123456779	22011966	1	P.O. Box 134

Deterministic record linkage involves an item-for-item computerized comparison where everything needs to match exactly, as shown in Fig. WC.1.1.

Record linkage is much easier if there is a unique personal record identifier that is common to the databases being compared. Where a single, common record identifier is not available, a combination of unique identifier variables can be used to allow the discrimination of records. For example, the combination of “name of patient”, “region” and “date of birth”.

In practice, there are often slight variations in the data contained in two files for the same variables; for instance, names may be misspelt, phone numbers may be mistyped and variables may be missing from one of the files under comparison, as shown in Fig. WC.1.2.

The variations in Fig. WC.1.2 would prevent a deterministic match from being made. Thus, deterministic matching might miss a significant number of true matches, necessitating an enormous amount of manual reviewing of results.

Probabilistic record linkage processes can accommodate some missing values, misspellings, abbreviations, typographical errors and other errors. Manual review involves using intuition to help in identifying true matches for records that contain slight variations in data between two files for the same variable. For example, in Fig. WC.1.2, because the categories “Last Name”, “Sex” and “Address” all match, and the typographical errors and transpositions involved are common, the two records would still be deemed a match. Probabilistic linkage software translates intuition into formal decision-making rules using the concept of most probable transpositions. Probabilistic record linkage estimates the probability that two records are for the same person, generating a score that indicates, for any pair of records, how likely it is that they both refer to the same person. The result is a sorted list of the likely and possible matched pairs, arranged in order of their scores. The total linkage score between any two records is the sum of the scores generated from matching individual fields.

Probabilistic record linkage is uncertain in nature and should only be used in the absence of unique, exact and reliable individual identifiers.

Chapter 2
Steps in the record linkage process

Record linkage comprises several steps: data cleaning and standardization, selection of matching variables, blocking or indexing, searching and scoring, and manual review (Fig. WC.2.1).

2.1 Data cleaning and standardization

Data cleaning and standardization is the first step in the record linkage process. The data to be used may be recorded in different formats, contain errors and inconsistencies, or have items missing. The aim of data cleaning is to convert the input data into a defined format and resolve all inconsistencies. Possible data transformations include the following:

- removal of:
    - commas and other punctuation marks; this is particularly important if the file has to be transformed into a delimited format (e.g. CSV), to be used by the linkage software;
    - leading, trailing or internal unnecessary white spaces and unseen characters (i.e. trimming of words);
    - numbers from variables that should be composed purely of letters, and vice versa;
    - accent marks;
- correction of variations of upper and lower case;
- standardization of:
    - terms that indicate lack of information (e.g. “don’t know”, “unknown”, “unidentified” or “n/a”);
    - date formats;
    - terms used in addresses (e.g. “St.” could be replaced by “Street” and “Sq.”by “Square”);
    - the order of the address elements; and
- for common words and names, replacement of obvious spelling variations with standard spelling.

2.2 Selection of matching variables

The variables selected to match records should be chosen on the basis of their suitability for discriminating between different records. A variable that has many different values has a higher discriminative power; for example, the discriminating power of a comparison between two different records containing the same last name is greater if that name is rare rather than common. Variables with a high proportion of missing values are not very useful for matching records. Because methods for probability matching depend on making comparisons between each of several variables with

Fig. WC.2.1 Key steps in the record linkage process

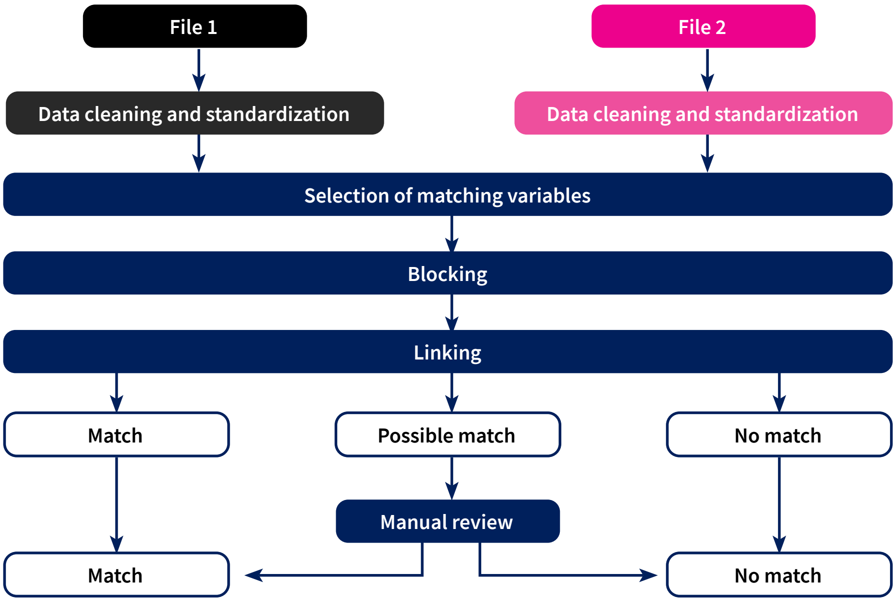
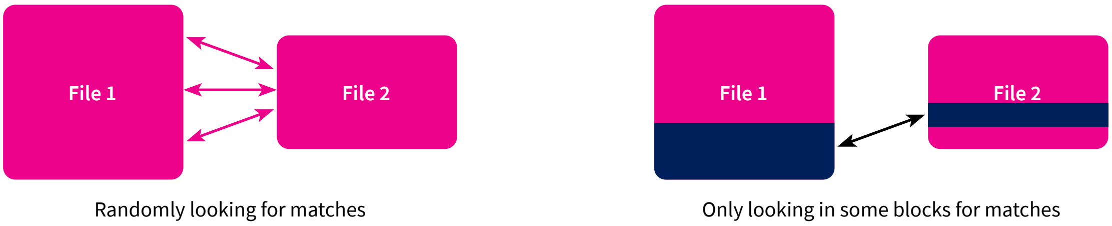

File 1
File 2
Data cleaning and standardization
Data cleaning and standardization
Selection of matching variables
Blocking
Linking
Match
Possible match
No match
Manual review
Match
No match

Fig. WC.2.2 Example of blocking


File 1
File 2
Randomly looking for matches
File 1
File 2
Only looking in some blocks for matches

identifying information, variables such as name, date of birth, name of mother and address are commonly used in combination. If the order of the elements in the address variable cannot be satisfactorily arranged in the preprocessing phase, it may be best not to choose "address" as a matching variable (although address may be used later as part of the manual verification of uncertain linked pairs).

2.3 Blocking

When comparing records in two files, in theory, each record in one file should be compared with every record in a second file. In practice, for files with large numbers of records, the total number of possible pairs is too large for reasonable computation. As the number of records to be compared increases in a linear fashion, the computational task increases quadratically. Blocking involves splitting the data sets into smaller groups or strata, then comparing only files in records within particular blocks, as shown in Fig. WC.2.2.

"Sex" may be a good blocking variable in the sense that not many records are likely to be wrongly classified or to have missing variables in this field. However, blocking by sex only splits the file into two parts, which would not impart great gains in the efficiency of searching. Choosing "district of residence" or "postcode" as a blocking variable will certainly have a higher impact in increasing the efficiency of searching. However, the records of patients who have moved from one area to another during the study period will not be paired. Phonetic codes for last name and year of birth have also been used as blocking variables.

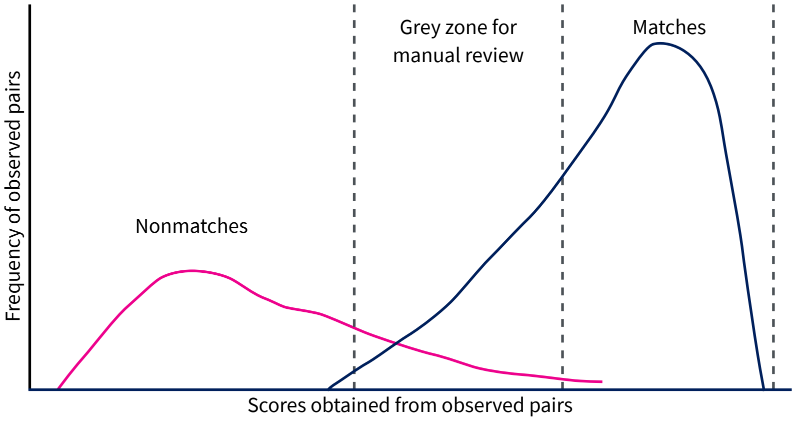
2.4 Searching and scoring

Searching and scoring are the core of the linkage process – they represent the phase in which the computer does the work. In the probabilistic linkage process, the computer program searches probable pairs of records, estimating the probability that the pairs relate to the same person, calculating linkage scores and displaying the results in an ordered way.

In a typical probabilistic record linkage, the histogram of the score frequency of observed pairs will show a bimodal distribution (Fig. WC.2.3). One mode is for pairs with low scores and another is for pairs with high scores. Pairs around and below the mode with low scores can usually be classified as nonmatches without a need for further review. The same is true for pairs around and above the mode with high scores, which can be directly classified as matches. The problem lies in the so-called grey zone between the two modes, where pairs need to be further reviewed to be satisfactorily classified.

Selecting score cutoff points that will separate matches from nonmatches requires experience with probabilistic linkage. The process tends to be a trade-off between being confident that all matches are correctly identified

Fig. WC.2.3 Distribution of scores obtained from observed pairs



and not leaving too much work to be done in the manual review process. It is preferable to start with the default cutoff values recommended by the software program selected.

2.5 Manual review

The final step in the record-linkage process is manual review; that is, the process of manually looking at uncertain linkages in the grey zone and then classifying them as matches or nonmatches. Manual review of data involves using intuition in combination with, for example, a knowledge of the frequency of names and addresses in the population, to help decide whether paired records relate to the same person even though they contain slight variations or are missing certain information. In theory, the person undertaking the review can access additional data from variables not used in the searching and scoring phase, or even data external to the files being compared, which enables the person to resolve the linkage status. Once all pairs have been classified, it is best to **not** delete the unwanted duplicated records. Instead, duplicated records should be marked as duplicates and kept on file, to allow a subsequent reassessment. Any record-linkage exercise should be accompanied by full documentation of the methods used. The documentation is necessary to allow for peer review and to provide a record of what has been done for possible replication in the future.

Chapter 3
Preparing for an in-country record linkage exercise

Owing to the importance and intensity of the activity, it is recommended that countries plan and allocate required resources before embarking on preparations for an in-country record linkage exercise. The resources or conditions necessary for a successful exercise are as follows:

- *System level preparation* – NTPs should conduct a mapping exercise of their routine programme information flows from service delivery points in facilities and in communities or workplaces. This exercise should include mapping of existing digital data registers that could inform the programme's epidemiological surveillance activities.
- *A verified digital version of the TB register* – NTPs should compile a master TB register in electronic format, containing information on the demographic identifiers of each notified case (e.g. patient identification number, national identification number where applicable, names, gender, date of birth and area of residence). The list will need to be verified to ensure that the variables captured are transcribed as accurately as possible.
- *Careful management of access to the registers used for linkage* – This is required to prevent inadvertent contravention of confidentiality or exposure of sensitive patient records. Therefore, records and linkage logs should not be printed unless absolutely necessary, and any printouts should be safely stored in line with country ethical guidance on handling patient data.
- *Computing resources* – To conduct probabilistic linkage, the NTP may need to decide which software packages to use. The annexes to this document describe a simple software for matching, Link Plus, but other software and algorithms may be used, depending on country preferences.
- *Human resources* – A multidepartmental team should be set up that includes members from the NTP M&E office, the civil registration and vital statistics office, the ministry of health (MoH) health informatics and standards office, the epidemiology unit, the laboratory services department, the planning unit, the community strategy unit and other relevant units, all under the guidance of the NTP manager.
- *Data required to inform the interpretation of the linkage results* – Such data include previous epidemiological review results and documentation, annual country health reports and national health accounts or statistics.

NTPs should plan to conduct the linkage exercise – from record verification and register cleanup, to identification of missed cases, and finally to review of programme activities – over 3–4 weeks. They should plan for multisectoral meetings, to draw out inclusive and joint plans to address the programme gaps that emerge from this activity. These plans would address the patient and information flow processes across both the public and the private sector, and the process of linkage of diagnosed patients to TB care programmes. It is therefore recommended that NTPs include in their annual budget a comprehensive activity plan on record linkage and identification of missed cases.

Annex 1

Instructions for conducting external record linkage using the Link Plus application

Adapted from Link Plus Stand-Alone Probabilistic Record Linkage Software, Linkage exercises, Oregon, May 2006

Link Plus is a probabilistic record linkage program developed at the Cancer Division of the United States Centers for Disease Control and Prevention (US CDC). The main advantage of Link Plus is that it allows partial (approximate) matches on character variables (patient names) and fuzzy matching using phonetic coding systems as well as other variable-specific matching methods.

This annex provides an exercise that outlines how to use Link Plus to conduct external record linkage. Using two simulated registers – a TB register and a laboratory register – it takes the user through the steps needed to import the data, link the files and review the records, to obtain a list of patients who have tuberculosis (TB) but are not being treated for it.

Getting started

- Download and install the Link Plus software (www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm).
- The files to be linked (tb_register and laboratory_register) can be downloaded from here (link 1 and link 2). Save these files in the subfolder RegPlus > LinkPlus > data folder as .csv files (i.e. comma delimited). Open Local Disk (D:)> RegPlus > LinkPlus > data

This linkage exercise involves linking a simulated TB register with a simulated laboratory register of patients with laboratory-confirmed TB.

The **Simulated TB register** is a comma-delimited .csv file containing 80 000 records and the following variables:

Id	consecutive number of records from 1 to 80 000
last_name	patient's last name
first_name	patient's name
middle_name	patient's middle name
sex	2=female, 1=male
dob	date of birth formatted YYYYM MDD, missing day 99, missing month 99, missing year 9999
snn	social security number (XXXXXXXXX), missing value 999999999
region	patient's area of residence

The **Simulated laboratory register with bacteriologically confirmed TB patients** is a comma-delimited .csv file containing 20 000 records with the following variables:

Id	consecutive number of records from 1 to 20 000
last_name	patient's last name
first_name	patient's name
middle_name	patient's middle name
sex	2=female, 1=male
dob	date of birth formatted YYYYM MDD, missing day 99, missing month 99, missing year 9999
snn	social security number (XXXXXXXXX), missing value 999999999
region	patient's area of residence

In this linkage exercise, File 1 is the TB register and File 2 is the laboratory register. At the completion of a linkage run, in addition to the linkage report, Link Plus generates a non-match report, with the default name **Non_MatchReport.txt**, and stores it in the Report folder of the Link Plus directory. The report is a tab-delimited text file and can be opened with a text editor or a spreadsheet program. The Non_MatchReport.txt report will contain records from File 2, which, when matched to records in File 1, receive a linkage score below the specified cutoff value (i.e. records in File 2 that are definitely not matches with any records in File 1). Because the aim is to identify the patients that have been diagnosed with TB but were not included in TB register, we will specify the laboratory register as File 2 and the TB register as File 1 (Fig. WC.A1.1). Fig. WC.A1.2, below, summarizes the 10 steps in the external linkage process.

Fig. WC.A1.1 Overview of input and output files used in the external record linkage process

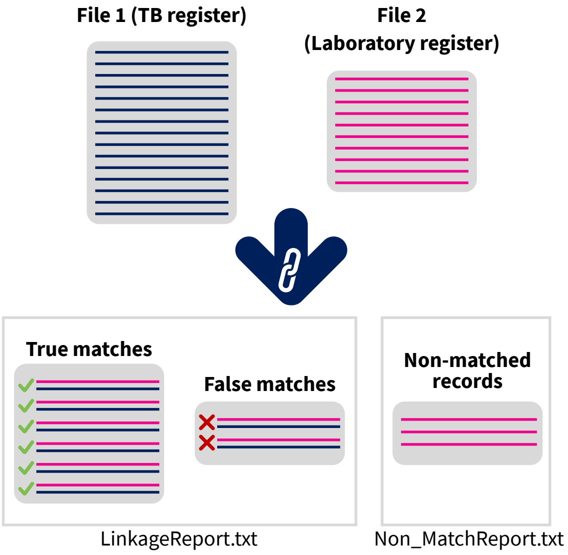
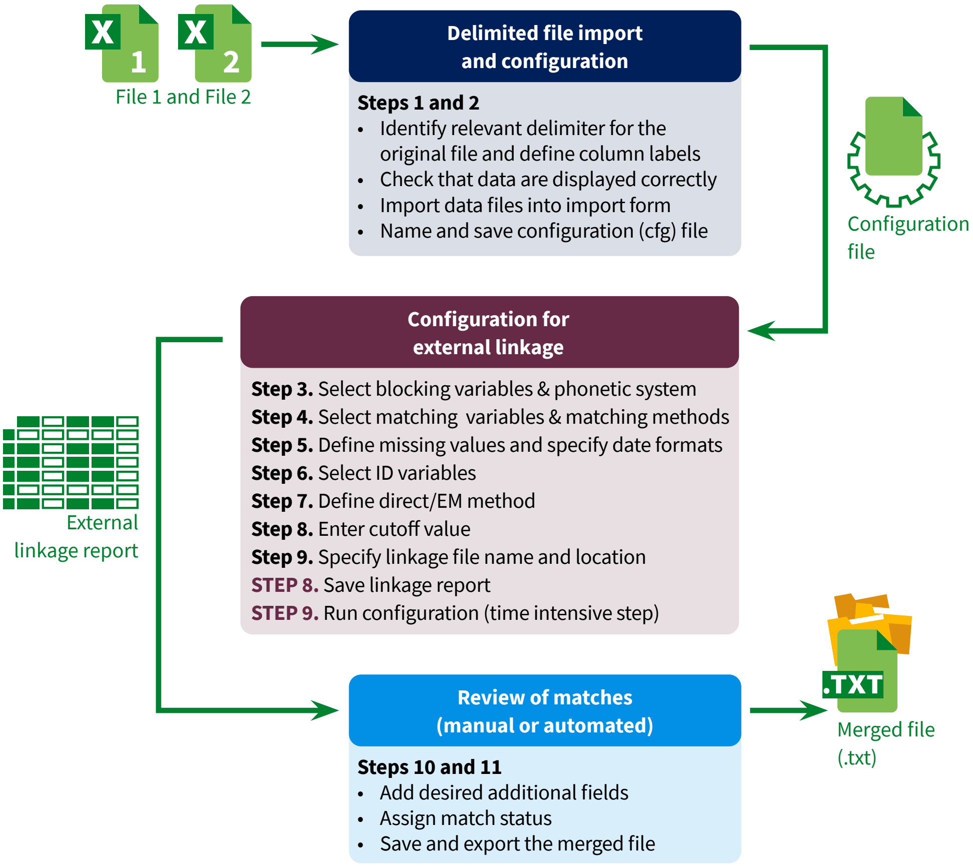


Fig. WC.A1.2 Overview of the steps in the external record linkage process using Link Plus



Delimited file import and configuration

Steps 1 and 2
- Identify relevant delimiter for the original file and define column labels
- Check that data are displayed correctly
- Import data files into import form
- Name and save configuration (cfg) file

Configuration for external linkage

Step 3. Select blocking variables & phonetic system
Step 4. Select matching variables & matching methods
Step 5. Define missing values and specify date formats
Step 6. Select ID variables
Step 7. Define direct/EM method
Step 8. Enter cutoff value
Step 9. Specify linkage file name and location
STEP 8. Save linkage report
STEP 9. Run configuration (time intensive step)

Review of matches (manual or automated)

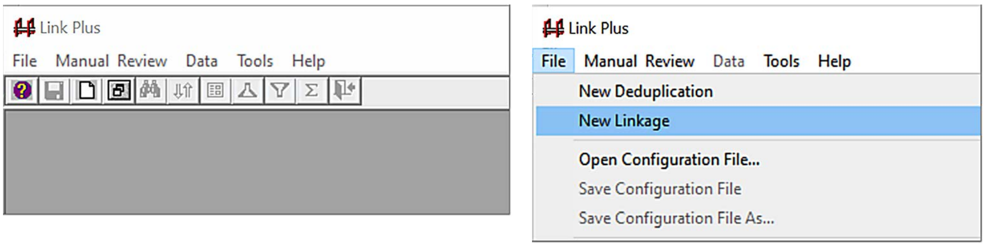
Steps 10 and 11
- Add desired additional fields
- Assign match status
- Save and export the merged file

Launching Link Plus and starting the external linkage exercise

Launch the Link Plus application by clicking on Start>All Programs>Registry Plus>Link Plus>Link Plus.

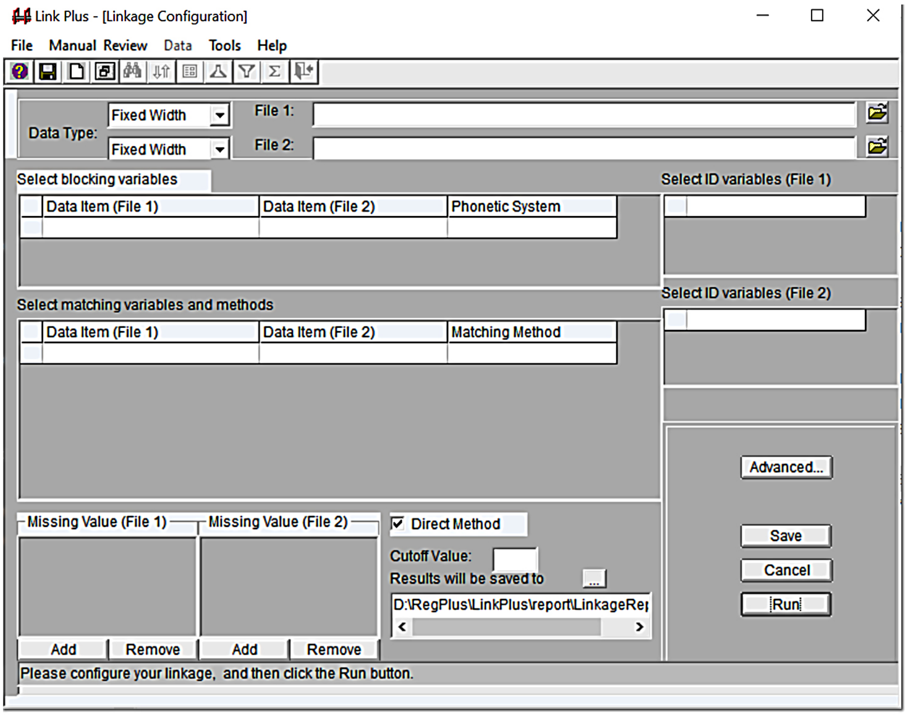
Result: The main Link Plus window opens (Fig. WC.A1.3); click on the **File** menu and select **New Linkage** (Fig. WC.A1.4).

Fig. WC.A1.3 Link Plus window


Link Plus
File Manual Review Data Tools Help
Link Plus
File Manual Review Data Tools Help
New Deduplication
New Linkage
Open Configuration File...
Save Configuration File
Save Configuration File As...

Result: The Link Plus Linkage Configuration window opens.

Fig. WC.A1.4 Link Plus linkage configuration window



Data Type: Fixed Width

Fixed Width

☑ Direct Method

D:\RegPlus\LinkPlus\report\LinkageRe

Please configure your linkage, and then click the Run button.

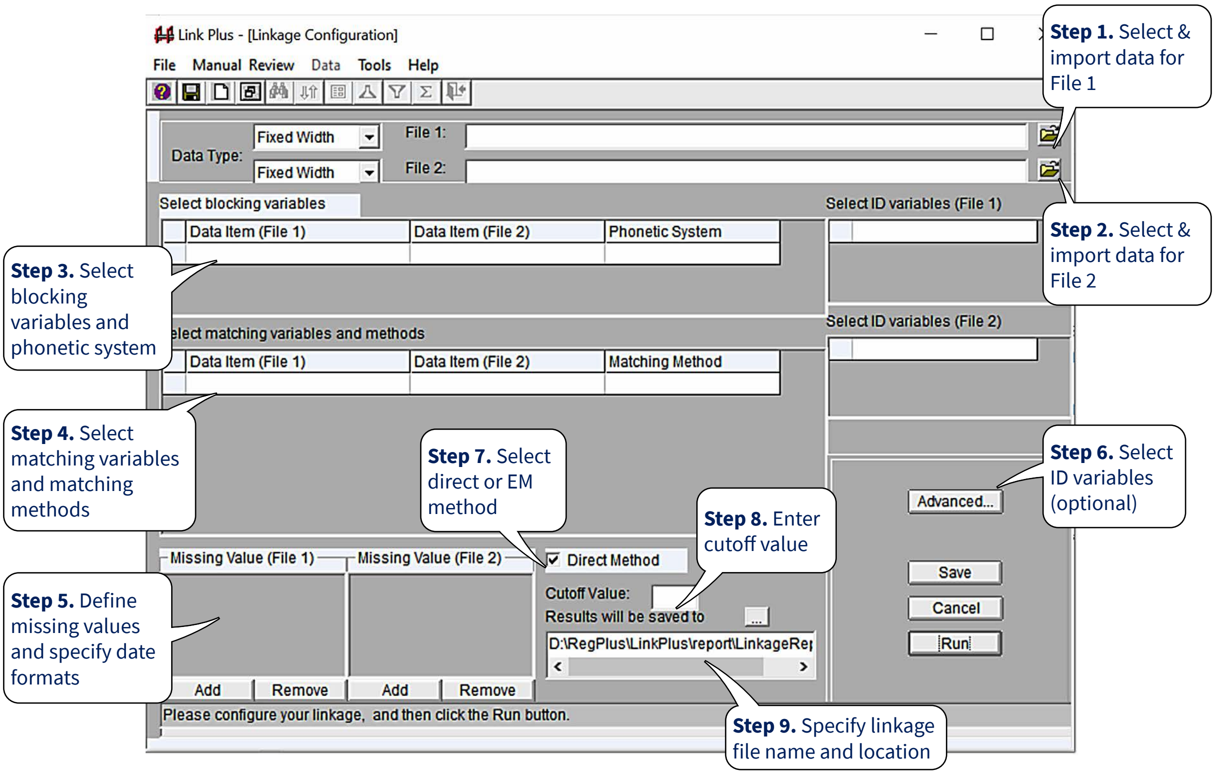
Review of external record linkage steps

There are 11 steps in the process of external linkage, summarized here and detailed below:

1. Select and import data for File 1
2. Select and import data for File 2
3. Select blocking variables and phonetic system
4. Select matching variables and matching methods
5. Define missing values and specify date formats
6. Select ID variables
7. (Optional) Select direct or EM method
8. Enter cutoff value
9. Specify linkage file name and location
10. Manually review uncertain matches
11. Export merged file

Steps 1–9 can be performed in the configuration window, as shown below (Fig. WC.A1.5). Steps 10 and 11 involve manual review of uncertain matches, and are performed on a different panel, which is described further below (Fig. WC.A1.27).

Fig. WC.A1.5 Overview of the first nine steps in the external record-linkage process



Step 1. Select & import data for File 1

Step 2. Select & import data for File 2

Step 3. Select blocking variables and phonetic system

Step 4. Select matching variables and matching methods

Step 5. Define missing values and specify date formats

Step 6. Select ID variables (optional)

Step 7. Select direct or EM method

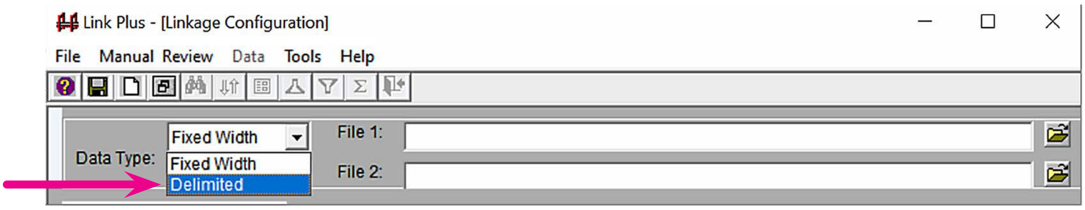
Step 8. Enter cutoff value

Step 9. Specify linkage file name and location

Step 1: Select and import data for File 1

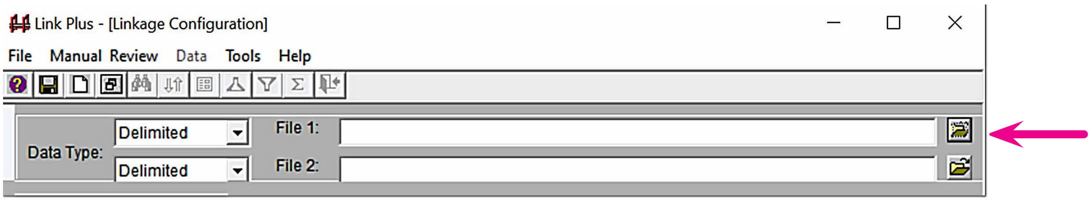
To select the data type for File 1, select the **Delimited** option in the **Data Type** drop-down menu for File 1 (Fig. WC.A1.6).

Fig. WC.A1.6 Selection of data type for File 1


Link Plus - [Linkage Configuration]
File Manual Review Data Tools Help
Data Type:
Fixed Width
Fixed Width
Delimited
File 1:
File 2:

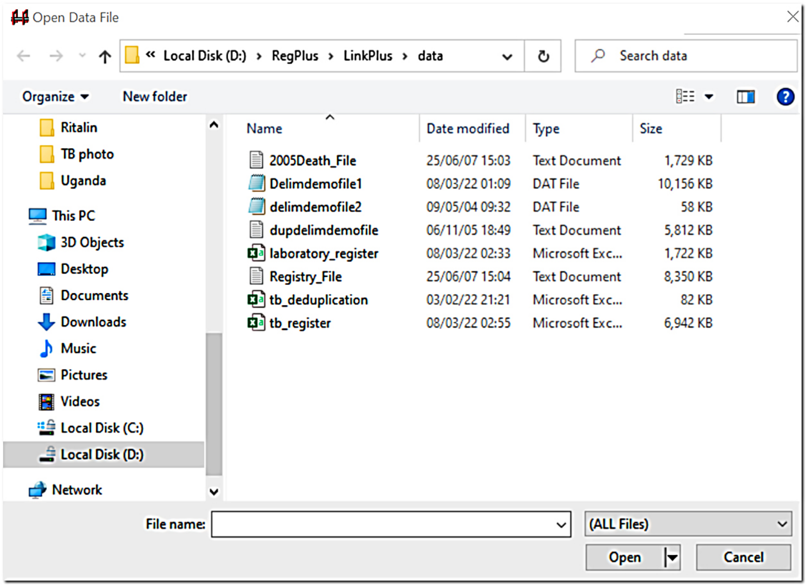
To locate the TB register data file and identify it as File 1, click on the folder icon to the right of the input box for File 1 (Fig. WC.A1.7).

Fig. WC.A1.7 Importing File 1


Link Plus - [Linkage Configuration]
File Manual Review Data Tools Help
Data Type:
Delimited
Delimited
File 1:
File 2:

Result: The Open Data File dialogue box opens (Fig. WC.A1.8).

Fig. WC.A1.8 Open Data File dialogue box



Open Data File

« Local Disk (D:) › RegPlus › LinkPlus › data

Search data

Organize ▾ New folder

- Ritalin
- TB photo
- Uganda
- This PC
- 3D Objects
- Desktop
- Documents
- Downloads
- Music
- Pictures
- Videos
- Local Disk (C:)
- Local Disk (D:)
- Network

Name	Date modified	Type	Size
2005Death_File	25/06/07 15:03	Text Document	1,729 KB
Delimdemofile1	08/03/22 01:09	DAT File	10,156 KB
delimdemofile2	09/05/04 09:32	DAT File	58 KB
dupdelimdemofile	06/11/05 18:49	Text Document	5,812 KB
laboratory_register	08/03/22 02:33	Microsoft Exc...	1,722 KB
Registry_File	25/06/07 15:04	Text Document	8,350 KB
tb_deduplication	03/02/22 21:21	Microsoft Exc...	82 KB
tb_register	08/03/22 02:55	Microsoft Exc...	6,942 KB

File name:

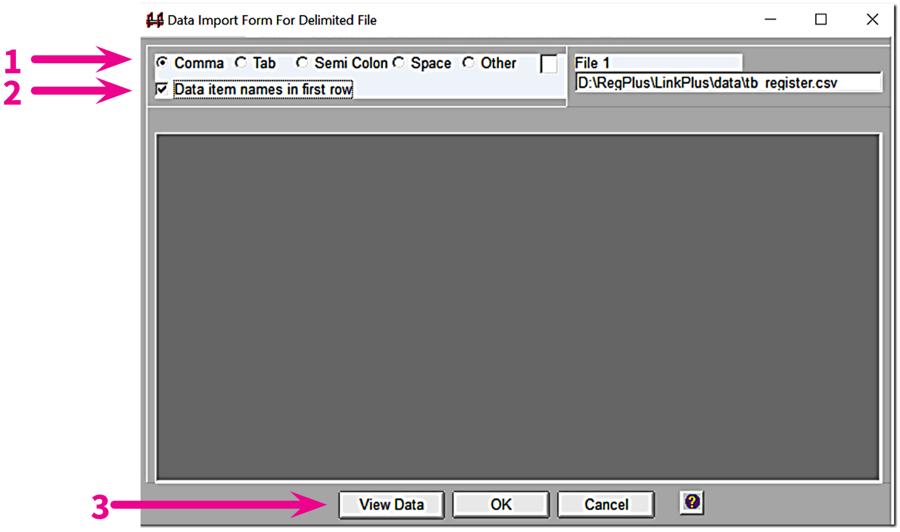
(ALL Files)

Open Cancel

Select the **tb_register.csv** file and click on **Open**.

Result: The Data Import Form For Delimited File dialogue box opens (Fig. WC.A1.9).

Fig. WC.A1.9 Data Import Form For Delimited File dialogue box



Data Import Form For Delimited File

● Comma ○ Tab ○ Semi Colon ○ Space ○ Other □

☑ Data item names in first row

File 1

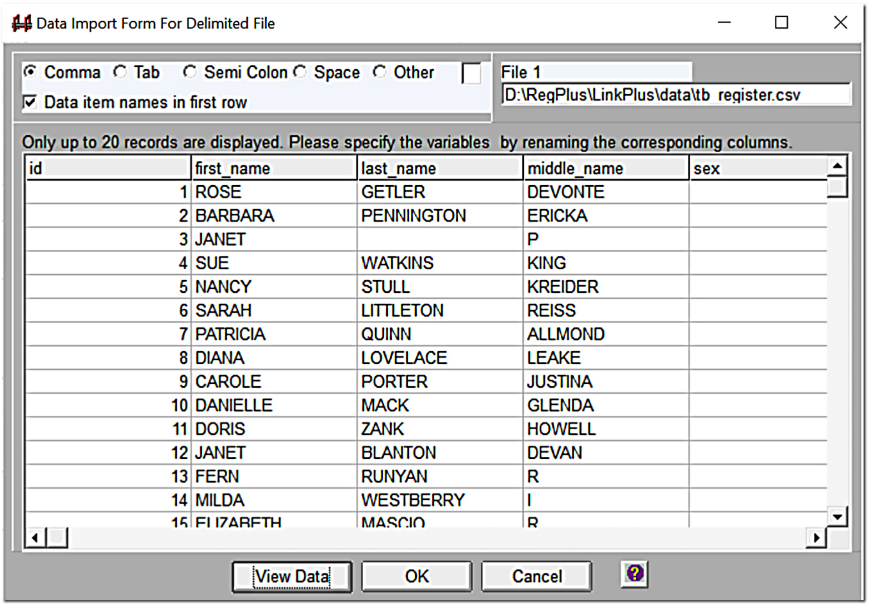
D:\RegPlus\LinkPlus\data\tb_register.csv

View Data OK Cancel ?

Select **Comma** as the delimiter type and tick the box next to **Data item names in first row**.

Click on **View Data** and make sure that variable names and values are well matched, then click **OK** (Fig. WC.A1.10). If date variables are present, check the format to ensure those variables have imported correctly, because Link Plus only recognizes date formats that are either in the formats YYYYMMDD or MMDDYYYY.

Fig. WC.A1.10 Data from File 1 imported into import form



Data Import Form For Delimited File

● Comma ○ Tab ○ Semi Colon ○ Space ○ Other □

☑ Data item names in first row

File 1

D:\RegPlus\LinkPlus\data\tb_register.csv

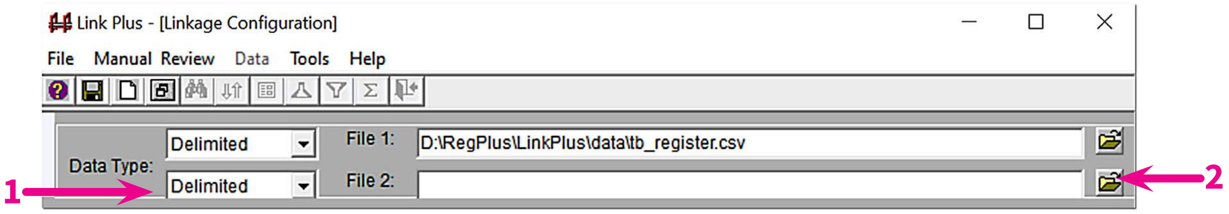
Only up to 20 records are displayed. Please specify the variables by renaming the corresponding columns.

id	first_name	last_name	middle_name	sex
1	ROSE	GETLER	DEVONTE	
2	BARBARA	PENNINGTON	ERICKA	
3	JANET		P	
4	SUE	WATKINS	KING	
5	NANCY	STULL	KREIDER	
6	SARAH	LITTLETON	REISS	
7	PATRICIA	QUINN	ALLMOND	
8	DIANA	LOVELACE	LEAKE	
9	CAROLE	PORTER	JUSTINA	
10	DANIELLE	MACK	GLENDA	
11	DORIS	ZANK	HOWELL	
12	JANET	BLANTON	DEVAN	
13	FERN	RUNYAN	R	
14	MILDA	WESTBERRY	I	
15	ELIZABETH	MASCIO	R	

Step 2: Select and import data for File 2

To import the laboratory register, first select the data type for File 2; that is, select the **Delimited** option in the **Data Type** drop-down menu for File 2. To locate the laboratory register data file and identify it as File 2, click on the folder icon to the right of the input box for File 2 (Fig. WC.A1.11).

Fig. WC.A1.11 Importing File 2



Link Plus - [Linkage Configuration]

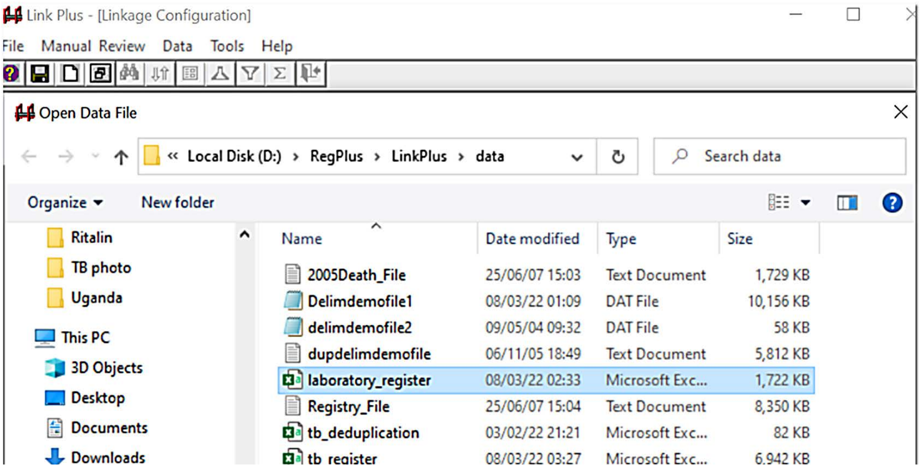
Data Type:

File 1: Delimited — D:\RegPlus\LinkPlus\data\tb_register.csv

File 2: Delimited —

Result: The Open Data File dialogue box opens (Fig. WC.A1.12).

Fig. WC.A1.12 Open Data File dialogue box



Link Plus - [Linkage Configuration]

Name	Date modified	Type	Size
2005Death_File	25/06/07 15:03	Text Document	1,729 KB
Delimdemofile1	08/03/22 01:09	DAT File	10,156 KB
delimdemofile2	09/05/04 09:32	DAT File	58 KB
dupdelimdemofile	06/11/05 18:49	Text Document	5,812 KB
laboratory_register	08/03/22 02:33	Microsoft Exc...	1,722 KB
Registry_File	25/06/07 15:04	Text Document	8,350 KB
tb_deduplication	03/02/22 21:21	Microsoft Exc...	82 KB
tb_register	08/03/22 03:27	Microsoft Exc...	6,942 KB

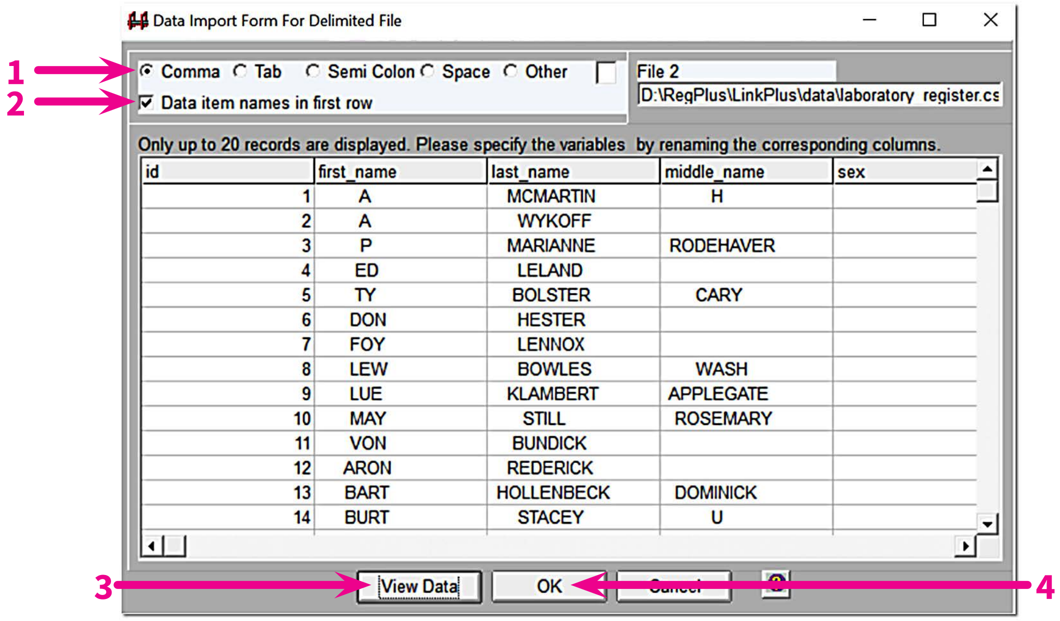
Select the **laboratory_register** file, and click on **Open**.

Result: The Data Import Form For Delimited File dialogue box opens.

Select **Comma** as the delimiter type and tick the box next to **Data item names in first row**.

Click on **View Data** and make sure that variable names and values are well matched, then click **OK** (Fig. WC.A1.13).

Fig. WC.A1.13 Data from File 2 imported into the import form



Data Import Form For Delimited File

● Comma ○ Tab ○ Semi Colon ○ Space ○ Other □

☑ Data item names in first row

File 2

D:\RegPlus\LinkPlus\data\laboratory_register.cs

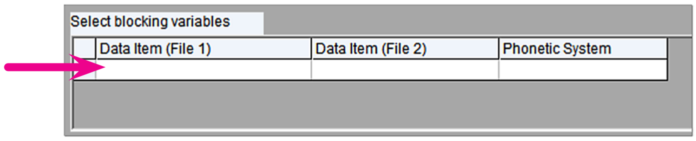
Only up to 20 records are displayed. Please specify the variables by renaming the corresponding columns.

id	first_name	last_name	middle_name	sex
1	A	MCMARTIN	H	
2	A	WYKOFF		
3	P	MARIANNE	RODEHAVER	
4	ED	LELAND		
5	TY	BOLSTER	CARY	
6	DON	HESTER		
7	FOY	LENNOX		
8	LEW	BOWLES	WASH	
9	LUE	KLAMBERT	APPLEGATE	
10	MAY	STILL	ROSEMARY	
11	VON	BUNDICK		
12	ARON	REDERICK		
13	BART	HOLLENBECK	DOMINICK	
14	BURT	STACEY	U	

Step 3: Select blocking variables and phonetic system

Click in the first empty cell of the grid in the **Select blocking variables** section of the Link Plus Linkage Configuration window (Fig. WC.A1.14).

Fig. WC.A1.14 Selection of blocking variables

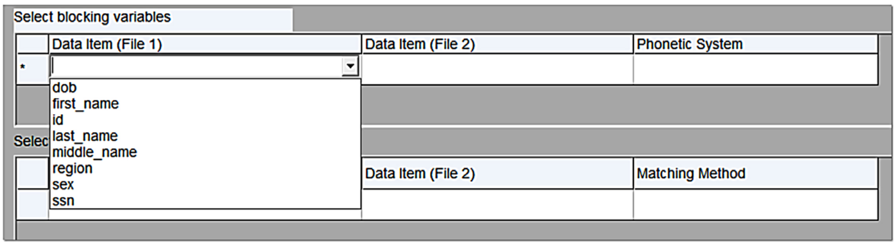


Select blocking variables

Data Item (File 1)	Data Item (File 2)	Phonetic System

Result: A drop-down menu appears that contains all of the variables that were specified for the simulated register in File 1 (**tb_register**) (Fig. WC.A1.15).

Fig. WC.A1.15 Selection of blocking variables (cont.)



Select blocking variables

Data Item (File 1)	Data Item (File 2)	Phonetic System

Using the drop-down menus for each of the three columns, make the selections shown in Table WC.A1.1 for blocking variables and phonetic systems for File 1 (**tb_register**) and File 2 (**laboratory_register**).

Table WC.A1.1 Selections for blocking variables and phonetic systems for Files 1 and 2

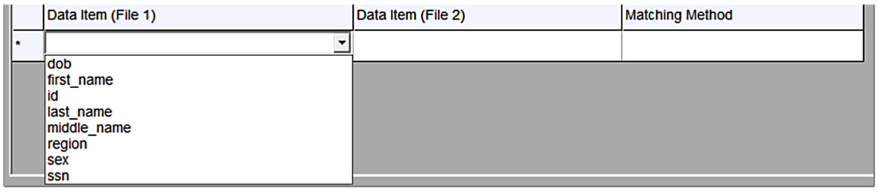
Data Item (File 1)	Data Item (File 2)	Phonetic System
first_name	first_name	NYSIIS
last_name	last_name	NYSIIS
dob	Dob	
ssn	ssn	

Step 4: Select matching variables and matching methods

Click on the first empty cell of the grid in the **Select matching variables and methods** section of the Link Plus Linkage Configuration window.

Result: A drop-down menu appears. This menu contains all the variables that were specified for the simulated register for File 1 (**tb_register**) (Fig. WC.A1.16).

Fig. WC.A1.16 Selection of matching variables



Using the drop-down menus for each of the three columns, make the selections shown in Table WC.A1.2 for matching variables and matching methods for Files 1 and 2.

Table WC.A1.2 Selections for matching variables and matching methods for Files 1 and 2

Data Item (File 1)	Data Item (File 2)	Matching Method
dob	dob	Date
last_name	last_name	Last Name
first_name	first_name	First Name
snn	Snn	SSN
middle_name	middle_name	Middle Name
sex	Sex	Exact

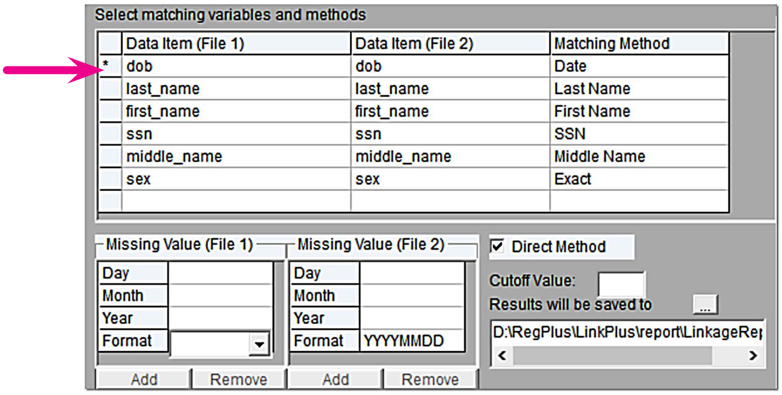
Step 5: Define missing values and specify date formats

Link Plus automatically treats null or empty values as missing data for matching variables; also, it allows the user to indicate additional values that are to be treated as missing data by the program.

To define missing values for the matching variables selected, first select the **date of birth** (dob) matching variable by clicking on its row in the matching variables grid.

Result: An asterisk appears on the grid row to indicate that the row is selected (see arrow below), and the Missing Value grids appear (Fig. WC.A1.17).

Fig. WC.A1.17 The Select matching variables and methods grid with date of birth selected



Select matching variables and methods

	Data Item (File 1)	Data Item (File 2)	Matching Method
*	dob	dob	Date
	last_name	last_name	Last Name
	first_name	first_name	First Name
	ssn	ssn	SSN
	middle_name	middle_name	Middle Name
	sex	sex	Exact

Missing Value (File 1)

Day	
Month	
Year	
Format	

Missing Value (File 2)

Day	
Month	
Year	
Format	YYYYMMDD

☑ Direct Method

Cutoff Value:

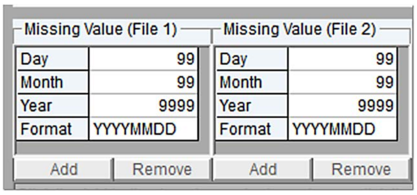
Results will be saved to

D:\RegPlus\LinkPlus\report\LinkageRe

Click separately in each cell of the missing values grid to enter the missing values, as shown in Fig. WC.A1.18.

The date format must also be specified. Link Plus currently accepts two date formats: MMDDYYYY and YYYYMMDD. Use the drop-down menu in the format row of the grid to select **MMDDYYYY** for both File 1 and File 2.

FFig. WC.A1.18 Specifying missing value formats for date of birth variable



Missing Value (File 1)

Day	99
Month	99
Year	9999
Format	YYYYMMDD

Missing Value (File 2)

Day	99
Month	99
Year	9999
Format	YYYYMMDD

Now do the following:

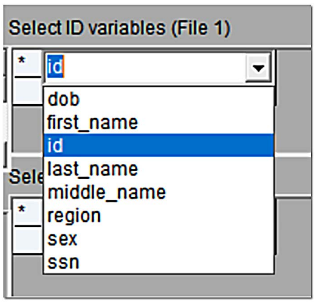
1. Select the SSN matching variable by clicking on its row in the matching variables grid.
2. Click on the **Add** button in the Missing Value grid for File 1.
3. Type 999999999 in the cell
4. Click on the **Add** button in the Missing Value grid for File 2 at the bottom of the screen.
5. Type 999999999 in the cell. You can add up to nine missing values.

Step 6: Select ID variables (optional)

Although the Link Plus linkage report and manual review screen will automatically list the matching variables that you have selected, you may want to define other variables to be included in the reports, called ID variables. Such variables (e.g. patient ID number) are used for identifying records. Selection of ID variables is optional.

Click in the Select ID variables (File 1) grid, and select the variable **id** from the dropdown list that appears, as shown in Fig. WC.A1.19, then do the same for File 2.

Fig. WC.A1.19 Selection of ID variables



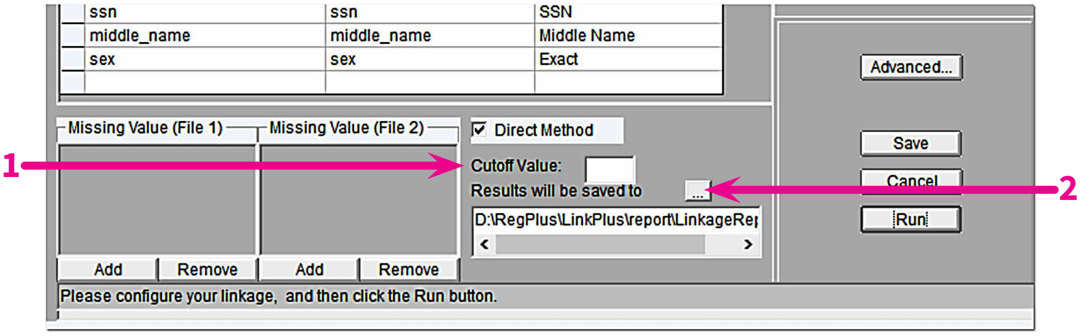
Select ID variables (File 1)

* id

Step 7: Select direct or EM method

For the purposes of the TB laboratory register linkage exercise, we will use the direct method to derive the M-probabilities used in the linkage. The M-probability is the probability that a matching variable agrees, given that a comparison pair is a match. When the Direct Method box is ticked, default M-probabilities are used in the linkage. Leave the Direct Method box ticked (Fig. WC.A1.20).

Fig. WC.A1.20 Selecting direct method and entering the cutoff value



	ssn	ssn	SSN
	middle_name	middle_name	Middle Name
	sex	sex	Exact

☑ Direct Method

Cutoff Value:

Results will be saved to

D:\RegPlus\LinkPlus\report\LinkageRe|

Please configure your linkage, and then click the Run button.

Step 8: Enter cutoff value

The cutoff value is the score value above which comparison pairs are accepted as potential links and presented for manual review. Enter an initial cutoff value of 7 in the Cutoff Value box (Fig. WC.A1.20).

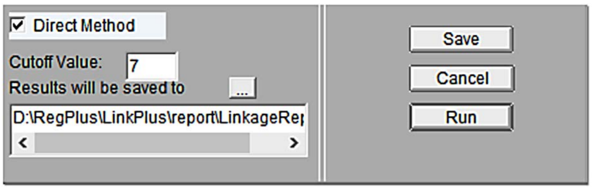
Step 9: Specify linkage file name and location

At the completion of a linkage run, Link Plus will generate a linkage report, named **LinkageReport.txt**, and will store it in the report folder of the Link Plus directory. The report is a tab-delimited text file, and it can be opened with the manual review feature of Link Plus (it can also be opened in a text editor or spreadsheet program). Records are presented in comparison pairs, sorted by their linkage scores in descending order. Pairs with scores above the selected cutoff value are listed. The row for each record contains all of the matching and ID variables used in the linkage.

Link Plus also generates a non-match report, named **Non_MatchReport.txt**, which contains records from File 2 not matched to records in File 1 (i.e. records receiving a linkage score below the specified cutoff value).

Save the linkage report under a new name, otherwise it will be overwritten the next time you run Link Plus. To rename the linkage report, first click on ... the to the right of **Results will be saved to** (Fig. WC.A1.21).

Fig. WC.A1.21 Renaming the linkage report file



☑ Direct Method

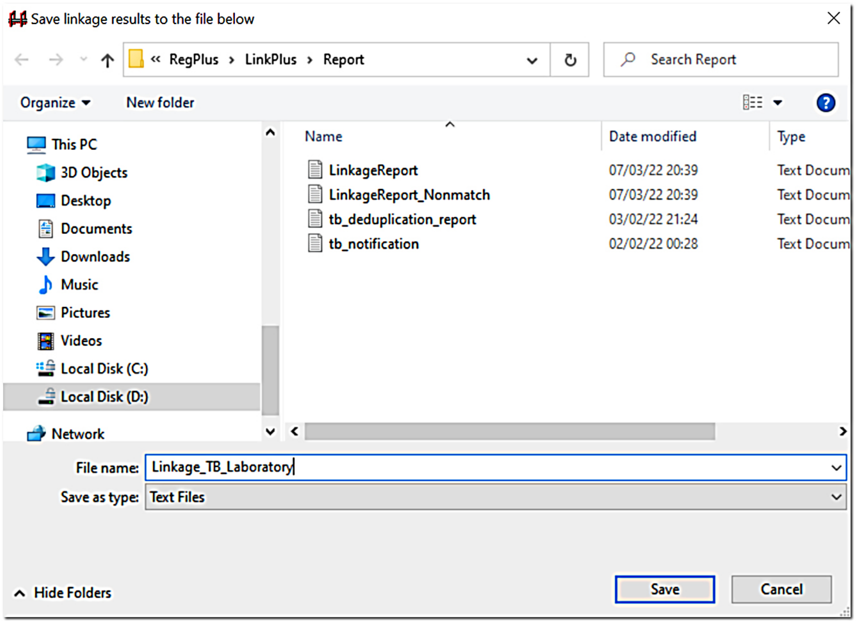
Cutoff Value: 7

Results will be saved to

D:\RegPlus\LinkPlus\report\LinkageRe|

Result: The **Save linkage results to the file below** dialogue box opens. Enter **Linkage_TB_Laboratory** in the File name box and click on **Save** (Fig. WC.A1.22**)**.

Fig. WC.A1.22 Dialogue box for saving linkage report file


Save linkage results to the file below

« RegPlus › LinkPlus › Report

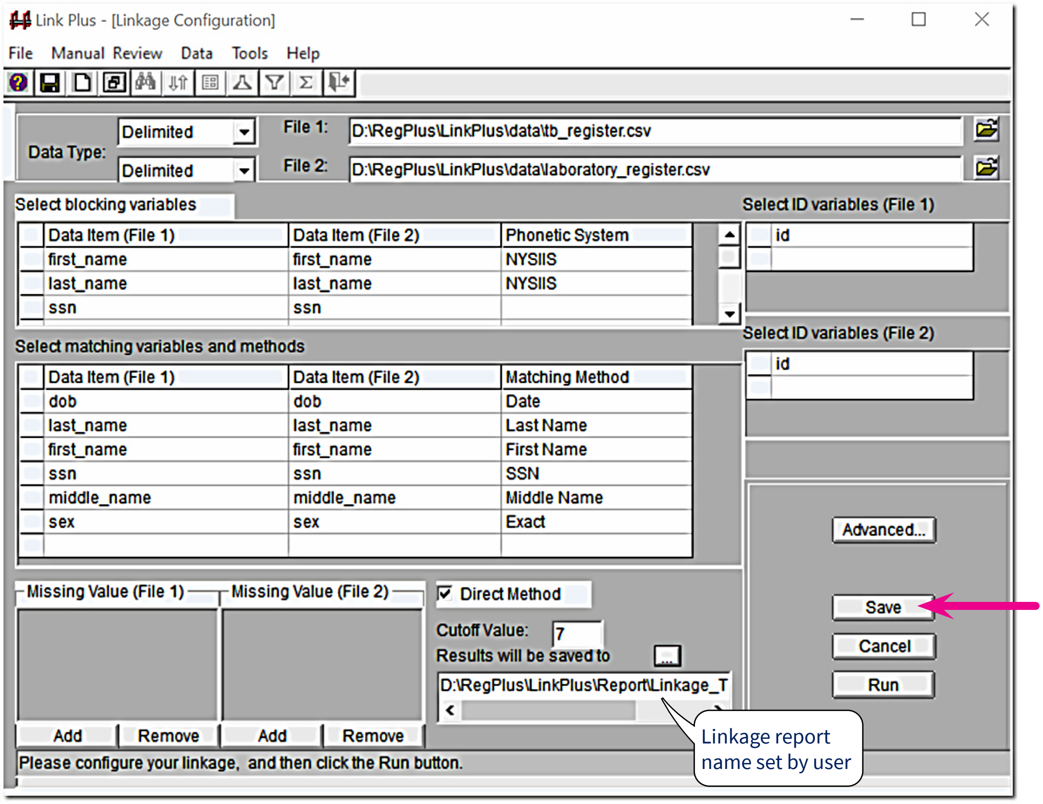
Name	Date modified	Type
LinkageReport	07/03/22 20:39	Text Docum
LinkageReport_Nonmatch	07/03/22 20:39	Text Docum
tb_deduplication_report	03/02/22 21:24	Text Docum
tb_notification	02/02/22 00:28	Text Docum

File name: Linkage_TB_Laboratory

Save as type: Text Files

Result: The application returns you to the Linkage Configuration window, where the linkage report file name has been changed to **Linkage_TB_Laboratory.txt** (Fig. WC.A1.23).

Fig. WC.A1.23 Linkage Configuration window showing renamed linkage report file


Link Plus - [Linkage Configuration]

Data Type: Delimited — File 1: D:\RegPlus\LinkPlus\data\tb_register.csv

Data Type: Delimited — File 2: D:\RegPlus\LinkPlus\data\laboratory_register.csv

Select blocking variables

Data Item (File 1)	Data Item (File 2)	Phonetic System
first_name	first_name	NYSIIS
last_name	last_name	NYSIIS
ssn	ssn	

Select ID variables (File 1): id

Select ID variables (File 2): id

Select matching variables and methods

Data Item (File 1)	Data Item (File 2)	Matching Method
dob	dob	Date
last_name	last_name	Last Name
first_name	first_name	First Name
ssn	ssn	SSN
middle_name	middle_name	Middle Name
sex	sex	Exact

☑ Direct Method

Cutoff Value: 7

Results will be saved to

D:\RegPlus\LinkPlus\Report\Linkage_T

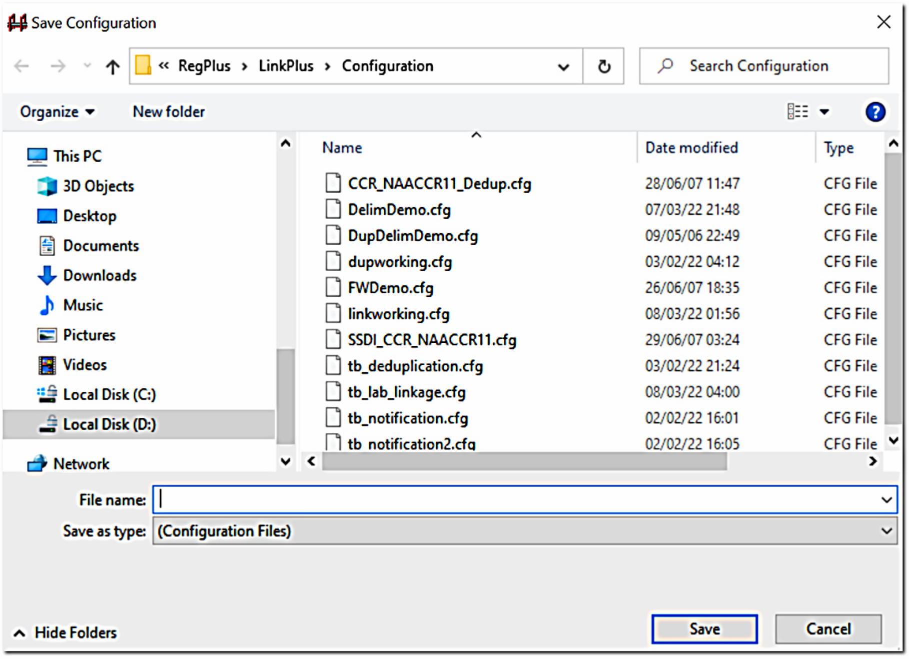
Please configure your linkage, and then click the Run button.

Linkage report name set by user

Click on **Save** to save the configuration file.

Result: The Save Configuration dialogue box opens (Fig. WC.A1.24).

Fig. WC.A1.24 Save Configuration dialogue box



Name	Date modified	Type
CCR_NAACCR11_Dedup.cfg	28/06/07 11:47	CFG File
DelimDemo.cfg	07/03/22 21:48	CFG File
DupDelimDemo.cfg	09/05/06 22:49	CFG File
dupworking.cfg	03/02/22 04:12	CFG File
FWDemo.cfg	26/06/07 18:35	CFG File
linkworking.cfg	08/03/22 01:56	CFG File
SSDI_CCR_NAACCR11.cfg	29/06/07 03:24	CFG File
tb_deduplication.cfg	03/02/22 21:24	CFG File
tb_lab_linkage.cfg	08/03/22 04:00	CFG File
tb_notification.cfg	02/02/22 16:01	CFG File
tb_notification2.cfg	02/02/22 16:05	CFG File

File name:

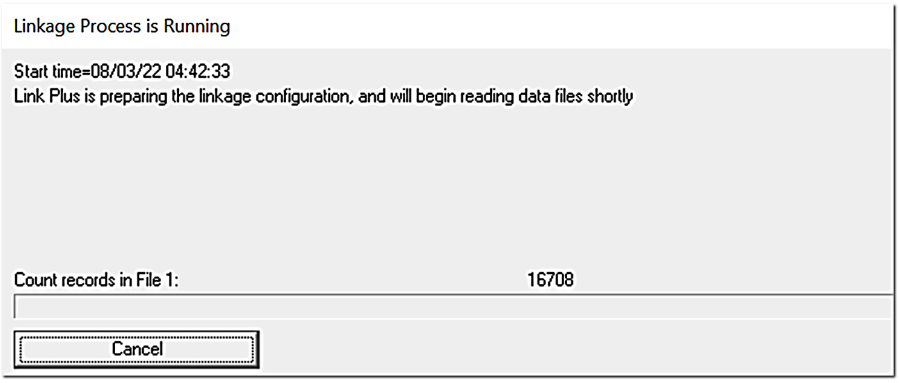
Save as type: (Configuration Files)

Enter **Linkage_TB_laboratory.cfg** in the File name box, and click on **Save**.

Click on **Run** to run the linkage process.

Result: The linkage begins and the **Linkage Process** progress window appears and provides the user with feedback about the linkage process as it is run (Fig. WC.A1.25).

Fig. WC.A1.25 Linkage Process Progress window (running)



Start time=08/03/22 04:42:33
Link Plus is preparing the linkage configuration, and will begin reading data files shortly

Count records in File 1: 16708

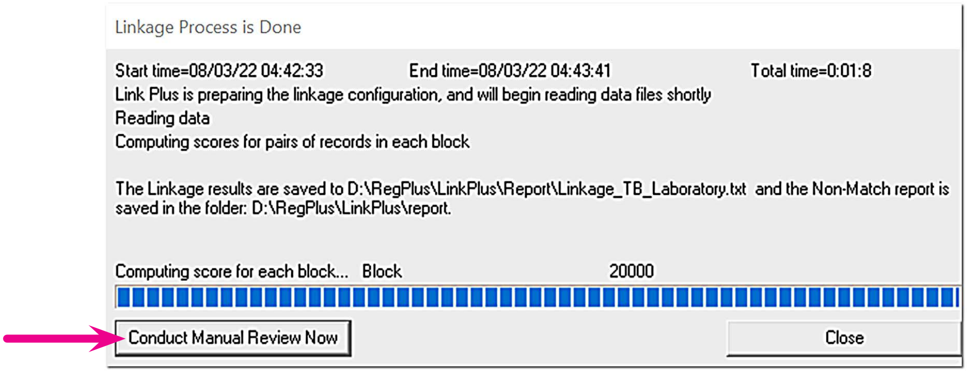
The progress window provides feedback about the preparation of the configuration, the reading of the data files, the blocking of the variables and the calculation of the linkage scores.

When the linkage process is complete, the progress window will present the user with the choice of conducting the manual review process directly after the linkage, or closing the progress window and conducting the manual review process at a later time.

Step 10: Manually review uncertain matches

One option for manually reviewing the potential matches generated by an external linkage (i.e. of File 1 with File 2) is to click on Conduct Manual Review Now in the Linkage Process progress window when the linkage has been completed (Fig. WC.A1.26).

Fig. WC.A1.26 Linkage process progress window (process ended)



Linkage Process is Done

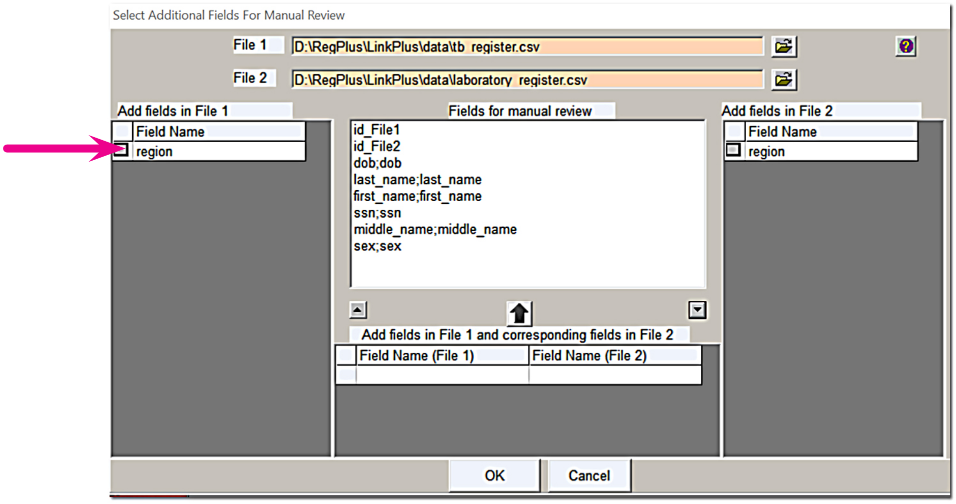
```
Start time=08/03/22 04:42:33          End time=08/03/22 04:43:41          Total time=0:01:8
Link Plus is preparing the linkage configuration, and will begin reading data files shortly
Reading data
Computing scores for pairs of records in each block

The Linkage results are saved to D:\RegPlus\LinkPlus\Report\Linkage_TB_Laboratory.txt  and the Non-Match report is
saved in the folder: D:\RegPlus\LinkPlus\report.

Computing score for each block...   Block                    20000
```

Another option is to open the linkage report by clicking on the **Manual Review** menu item in the Linkage Configuration window, then on the **New View** option. Either of these options will open the linkage report, where you can perform a manual review. This involves first selecting additional variables for manual review (if desired), then assigning a match status to each potential comparison pair.

Click on **Conduct Manual Review Now**.

Result: The **Select Additional Fields for Manual Review** dialogue box opens (Fig. WC.A1.27).

Fig. WC.A1.27 Select Additional Fields for Manual Review dialogue box



Select Additional Fields For Manual Review

File 1 D:\RegPlus\LinkPlus\data\tb_register.csv

File 2 D:\RegPlus\LinkPlus\data\laboratory_register.csv

Add fields in File 1

	Field Name
□	region

Fields for manual review

id_File1
id_File2
dob;dob
last_name;last_name
first_name;first_name
ssn;ssn
middle_name;middle_name
sex;sex

Add fields in File 1 and corresponding fields in File 2

Field Name (File 1)	Field Name (File 2)

Add fields in File 2

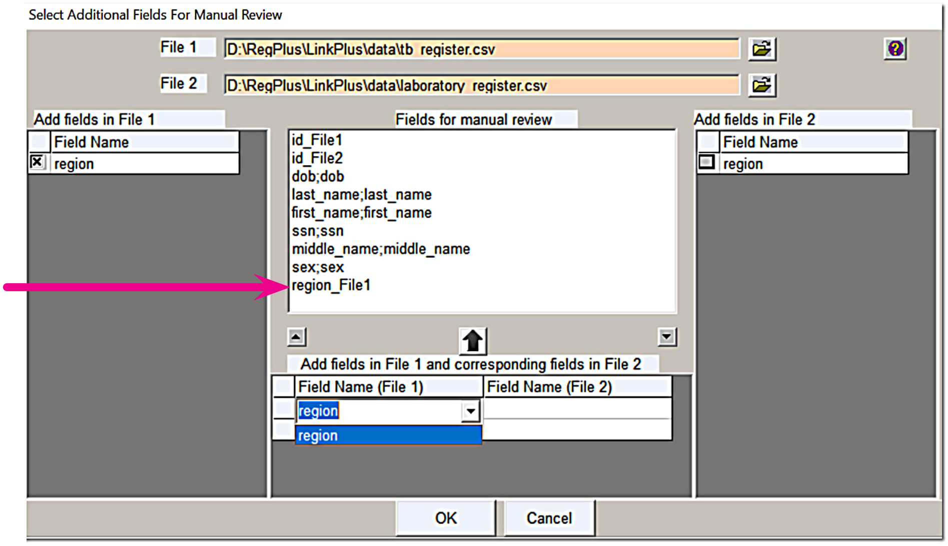
	Field Name
□	region

When manually reviewing potential matches, if you would like to view additional fields for the linkage (i.e. fields other than those already chosen for blocking, matching and ID purposes), you can tick the boxes on the left or right grids (or both) for the fields you would like to view. Fields added in this manner will be shown in separate columns on the manual review screen.

Click on the box to the left of **region** under Field Name in the **Add fields in File 1** grid.

Result: The region variable gets added to the central list of Fields for manual review, with a **_File1** after the field name to indicate that this field comes from File 1 (Fig. WC.A1.28).

Fig. WC.A1.28 Additional fields from File 1 added for manual review



Select Additional Fields For Manual Review

File 1 D:\RegPlus\LinkPlus\data\tb_register.csv

File 2 D:\RegPlus\LinkPlus\data\laboratory_register.csv

Add fields in File 1

	Field Name
☒	region

Fields for manual review

```
id_File1
id_File2
dob;dob
last_name;last_name
first_name;first_name
ssn;ssn
middle_name;middle_name
sex;sex
region_File1
```

Add fields in File 2

	Field Name
□	region

Add fields in File 1 and corresponding fields in File 2

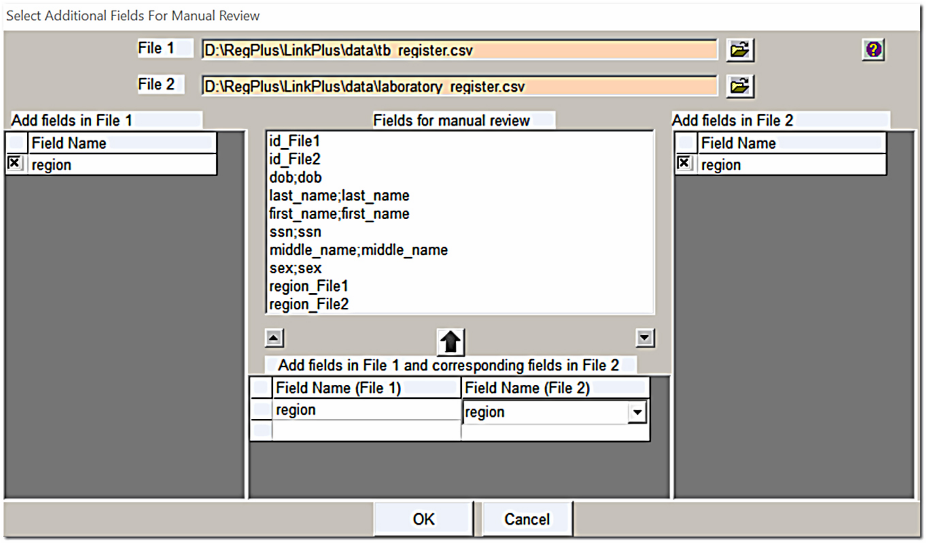
Field Name (File 1)	Field Name (File 2)
region	

To add the region variable from the File 2, click on the box to the left of **region** under Field Name in the **Add fields in File 2** grid. In the central panel, in the Field Name (File 2) grid, click in the first empty cell. Select **region** from the drop-down menu for File 2.

Click on the ⬆ button above the grids.

Result: The **region** field is added for manual review with the fields from each file to be displayed in the same column for each potential match (Fig. WC.A1.29).

Fig. WC.A1.29 Additional fields from File 1 and File 2 added for manual review



Select Additional Fields For Manual Review

File 1 D:\RegPlus\LinkPlus\data\tb_register.csv

File 2 D:\RegPlus\LinkPlus\data\laboratory_register.csv

Add fields in File 1

	Field Name
☒	region

Fields for manual review

```
id_File1
id_File2
dob;dob
last_name;last_name
first_name;first_name
ssn;ssn
middle_name;middle_name
sex;sex
region_File1
region_File2
```

Add fields in File 2

	Field Name
☒	region

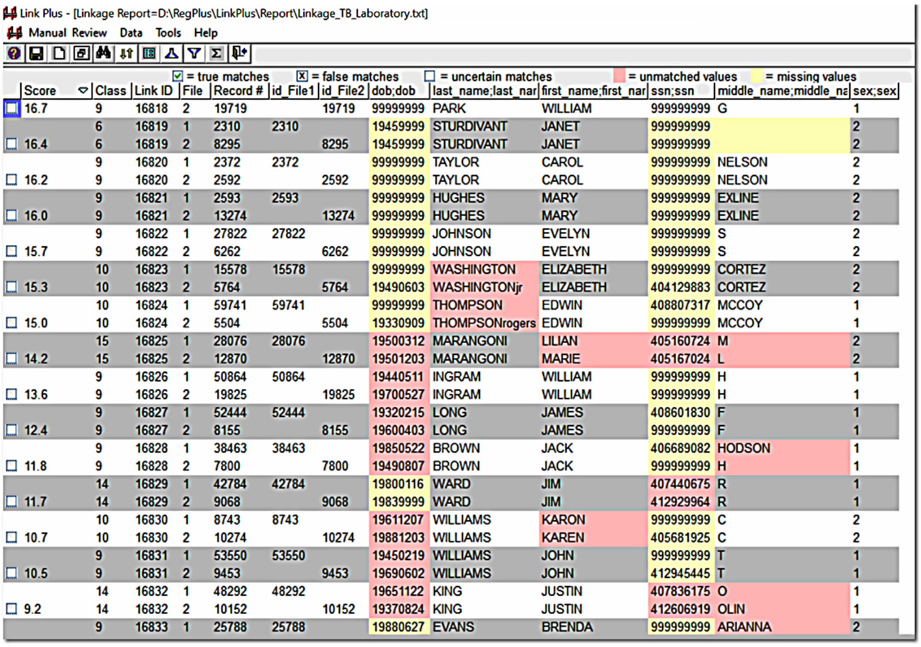
Add fields in File 1 and corresponding fields in File 2

Field Name (File 1)	Field Name (File 2)
region	region

Click on **OK**.

Result: The manual review screen opens for the **Linkage_TB_laboratory.txt** linkage report, with the fields and columns ordered as you specified (Fig. WC.A1.30).

Fig. WC.A1.30 Manual review screen



Link Plus - [Linkage Report=D:\RegPlus\LinkPlus\Report\Linkage_TB_Laboratory.txt]

Manual Review   Data   Tools   Help

☑ = true matches   ☒ = false matches   □ = uncertain matches   🟥 = unmatched values   🟨 = missing values

	Score	Class	Link ID	File	Record #	id_File1	id_File2	dob;dob	last_name;last_nar	first_name;first_nar	ssn;ssn	middle_name;middle_na	sex;sex
□	16.7	9	16818	2	19719		19719	99999999	PARK	WILLIAM	999999999	G	1
		6	16819	1	2310	2310		19459999	STURDIVANT	JANET	999999999		2
□	16.4	6	16819	2	8295		8295	19459999	STURDIVANT	JANET	999999999		2
		9	16820	1	2372	2372		99999999	TAYLOR	CAROL	999999999	NELSON	2
□	16.2	9	16820	2	2592		2592	99999999	TAYLOR	CAROL	999999999	NELSON	2
		9	16821	1	2593	2593		99999999	HUGHES	MARY	999999999	EXLINE	2
□	16.0	9	16821	2	13274		13274	99999999	HUGHES	MARY	999999999	EXLINE	2
		9	16822	1	27822	27822		99999999	JOHNSON	EVELYN	999999999	S	2
□	15.7	9	16822	2	6262		6262	99999999	JOHNSON	EVELYN	999999999	S	2
		10	16823	1	15578	15578		99999999	WASHINGTON	ELIZABETH	999999999	CORTEZ	2
□	15.3	10	16823	2	5764		5764	19490603	WASHINGTONjr	ELIZABETH	404129883	CORTEZ	2
		10	16824	1	59741	59741		99999999	THOMPSON	EDWIN	408807317	MCCOY	1
□	15.0	10	16824	2	5504		5504	19330909	THOMPSONrogers	EDWIN	999999999	MCCOY	1
		15	16825	1	28076	28076		19500312	MARANGONI	LILIAN	405160724	M	2
□	14.2	15	16825	2	12870		12870	19501203	MARANGONI	MARIE	405167024	L	2
		9	16826	1	50864	50864		19440511	INGRAM	WILLIAM	999999999	H	1
□	13.6	9	16826	2	19825		19825	19700527	INGRAM	WILLIAM	999999999	H	1
		9	16827	1	52444	52444		19320215	LONG	JAMES	408601830	F	1
□	12.4	9	16827	2	8155		8155	19600403	LONG	JAMES	999999999	F	1
		9	16828	1	38463	38463		19850522	BROWN	JACK	406689082	HODSON	1
□	11.8	9	16828	2	7800		7800	19490807	BROWN	JACK	999999999	H	1
		14	16829	1	42784	42784		19800116	WARD	JIM	407440675	R	1
□	11.7	14	16829	2	9068		9068	19839999	WARD	JIM	412929964	R	1
		10	16830	1	8743	8743		19611207	WILLIAMS	KARON	999999999	C	2
□	10.7	10	16830	2	10274		10274	19881203	WILLIAMS	KAREN	405681925	C	2
		9	16831	1	53550	53550		19450219	WILLIAMS	JOHN	999999999	T	1
□	10.5	9	16831	2	9453		9453	19690602	WILLIAMS	JOHN	412945445	T	1
		14	16832	1	48292	48292		19651122	KING	JUSTIN	407836175	O	1
□	9.2	14	16832	2	10152		10152	19370824	KING	JUSTIN	412606919	OLIN	1
		9	16833	1	25788	25788		19880627	EVANS	BRENDA	999999999	ARIANNA	2

Within comparison pairs, the record from File 1 is listed first, and the column headers for the variables selected for matching are displayed in a File 1;File 2 format. In addition, unmatched and missing values for matching variables are highlighted in pink and yellow, respectively.

Manual review screen menu items

From the main manual review screen, you can access the menu items for the Link Plus manual review process. Click on a menu item to perform the specified function. Table WC.A1.3 displays the menu items on the main manual review screen.

Table WC.A1.3   Menu items on manual review screen

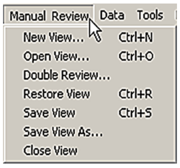
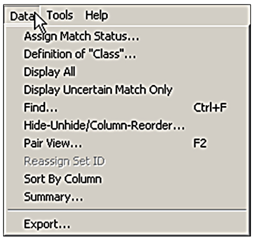
Menu	Menu item	Toolbar Icon	Function
Manual Review  Manual Review Data Tools New View... Ctrl+N Open View... Ctrl+O Double Review... Restore View Ctrl+R Save View Ctrl+S Save View As... Close View	New View...	🗋	Create a view from a linkage report
	Open View...		Open an existing view
	Double Review...		Create a view from two views to resolve any differences in assigned match status
	Restore View	▦	Open the most recent view
	Save View	💾	Save the current view
	Save View As...		Save the current view under a new name
	Close View...	🚪	Close the view
Data  Data Tools Help Assign Match Status... Definition of "Class"... Display All Display Uncertain Match Only Find... Ctrl+F Hide-Unhide/Column-Reorder... Pair View... F2 Reassign Set ID Sort By Column Summary... Export...	Assign Match Status...		Assign match status based on score
	Definition of "Class"...		Open the form where coding of the "class" category can be viewed
	Display All	⚗	Display all pairs for review, regardless of match status
	Display Uncertain Match Only	⏷	Hide all pairs that have been assigned a match status of true or false, displaying only uncertain matches for review

Table WC.A1.3 Menu items on manual review screen (continued)

Menu	Menu item	Toolbar Icon	Function
	Find	🔍	Find a word within a column
	Hide-Unhide/Column-Reorder		Hide and unhide columns and change the order of columns
	Pair View…	☰	View the current comparison pair in Pair View mode rather than Datasheet mode
	Reassign Set ID		Reassign Set ID for deduplication linkages only Set ID is an unique number assigned to each group of duplicated records and may be manually reassigned
	Sort By Column	⇅	Sort the data displayed on the manual review screen in ascending or descending order
	Summary…	Σ	Display summary information about the total number of true matches, false matches and uncertain matches for the current review session
	Export		Open the Merged File export dialogue box
Tools	Options…		**Manual Review tab:** Specify the colour scheme for manual review screen **Export tab:** Specify the merged export file delimiter, and exported match status **Data Link tab:** Specify the CRS Plus User
Help	Contents…	?	Open the Link Plus online Help

The datasheet view

For the manual review of uncertain matches, the default view for comparison pairs is the datasheet view (see Fig. WC.A1.30), where comparison pairs are sorted together and displayed in alternating rows of white and grey coloured background. For each comparison pair, in addition to the variables chosen for matching and manual review, Link Plus generates the outputs shown in Table WC.A1.4.

Table WC.A1.4 Outputs generated by Link Plus

Score or identifier	Description
Score	For a comparison pair, this is the overall weight across all matching variables – a higher score means a higher likelihood of being a match
Class	Each comparison pair is assigned to a "class" category, which is defined by the matching variables on which the pair matches exactly (e.g. Class 1 by default is defined by a perfect match on SSN, date of birth, and first and last names)
Link ID	A unique number assigned to each linked pair
File	For a comparison pair, the File number identifies the file from which (File 1 or File 2) the record in the comparison pair originated
Set ID	This is for deduplication linkages only – Set ID is a unique number assigned to each group of duplicated records because a record may have multiple duplicates; Set ID may be reassigned after assignment of match status has been completed
Record #	The record number is the sequential line (row) number for a record in its data file (e.g. the Record # of the record in the first line of a data file is 1)

When a linkage process is run, Link Plus saves all potential matches that generated a score above the specified cutoff value in the linkage report. When that report is opened for manual review, all potential matches will be presented for review, even those that generated high scores and are almost certain to be matches, and (depending on the cutoff value specified) those that generated low scores and are almost certain to be false matches. With practice, after initial review of the list of comparison pairs, it is often easy to identify an upper cutoff score above which all pairs are true matches, and a lower cutoff score below which all pairs are false matches. The remaining pairs are said to be in the “grey area”; that is, they are uncertain matches requiring manual review.

Manually assigning match status

Using one’s intuition in combination with, for example, a knowledge of local record keeping practices, one can manually assign a match status while viewing the manual review screen (see Fig. WC.A1.30). This can be done either by using the mouse to click on the **Match Status** option button next to the score column or by using the keyboard as described below.

Assigning true match status: For the comparison pair being reviewed, place the cursor in the **Match Status** tick box and left-click the mouse **once**, or press the **M** key to select the current pair as a true match.

Assigning false match status: For the comparison pair being reviewed, place the cursor in the **Match Status** tick box and left-click the mouse **twice**, or press the **N** key to select the current pair as a false match.

Assigning uncertain match status: For the comparison pair being reviewed, place the cursor in the Match Status tick box and left-click the mouse **three** times, or press the **B** key to select the current pair as an uncertain match.

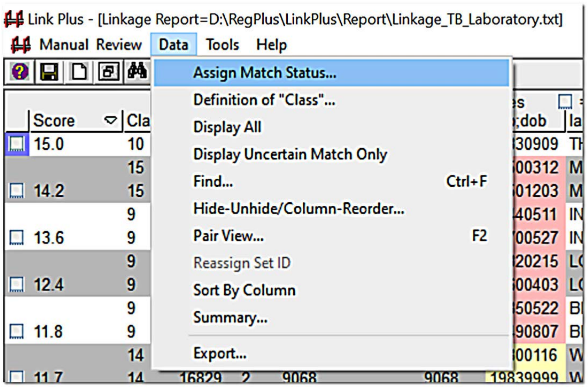
Assigning match status automatically by score

Rather than having to assign the match status to each comparison pair manually, Link Plus offers the option of automatically assigning match status for groups of pairs by their linkage score. For the current exercise, we will choose 14.2 as the upper cutoff limit and 8.9 as a lower cutoff limit

The steps for assigning match status automatically by score are outlined below.

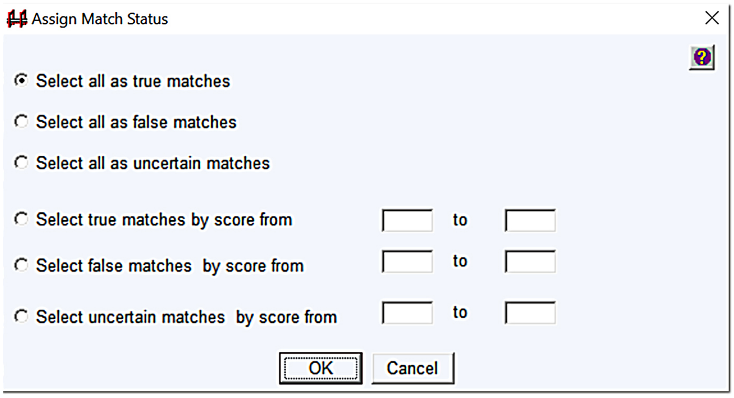
First, click on the **Data** menu, and select the **Assign Match Status…** option (Fig. WC.A1.31).

Fig. WC.A1.31 Opening the Assign Match Status dialogue box



Result: The Assign Match Status dialogue box opens (Fig. WC.A1.32).

Fig. WC.A1.32 Assign Match Status dialogue box



Assign Match Status

● Select all as true matches

○ Select all as false matches

○ Select all as uncertain matches

○ Select true matches by score from [ ] to [ ]

○ Select false matches by score from [ ] to [ ]

○ Select uncertain matches by score from [ ] to [ ]

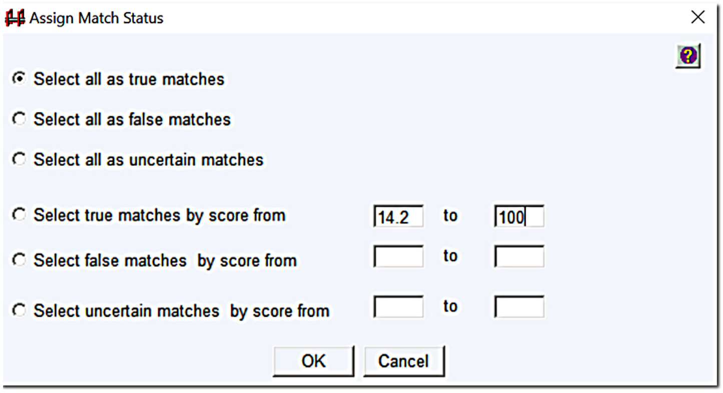
The Assign Match Status dialogue box offers the options shown in Table WC.A1.5.

Table WC.A1.5 Options in the Assign Match Status dialogue box

Option	Description
Select all as true matches	Selecting this option will assign a true match status to all comparison pairs in the current view
Select all as false matches	Selecting this option will assign a false match status to all comparison pairs in the current view
Select all as uncertain matches	Selecting this option will assign an uncertain match status to all comparison pairs in the current view
Select true matches by score	Selecting this option will assign a true match status to all comparison pairs that have a score in the range specified in the **from** and **to** input boxes
Select false matches by score	Selecting this option will assign a false match status to all comparison pairs that have a score in the range specified in the **from** and **to** input boxes
Select uncertain matches by score	Selecting this option will assign an uncertain match status to all comparison pairs that have a score in the range specified in the **from** and **to** input boxes

Click next to **Select true matches by score from**, enter **14.2** in the first box, enter **100** in the second box, and then click on **OK** (Fig. WC.A1.33).

Fig. WC.A1.33 Assigning cutoff of true matches



Assign Match Status

● Select all as true matches

○ Select all as false matches

○ Select all as uncertain matches

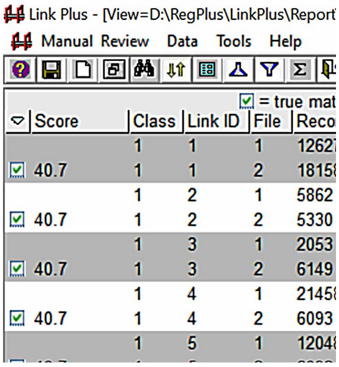
○ Select true matches by score from 14.2 to 100

○ Select false matches by score from ____ to ____

○ Select uncertain matches by score from ____ to ____

Result: The manual review screen reopens with the specified match statuses assigned (Fig. WC.A1.34).

Fig. WC.A1.34 Manual review screen with assigned match statuses



Link Plus - [View=D:\RegPlus\LinkPlus\Report

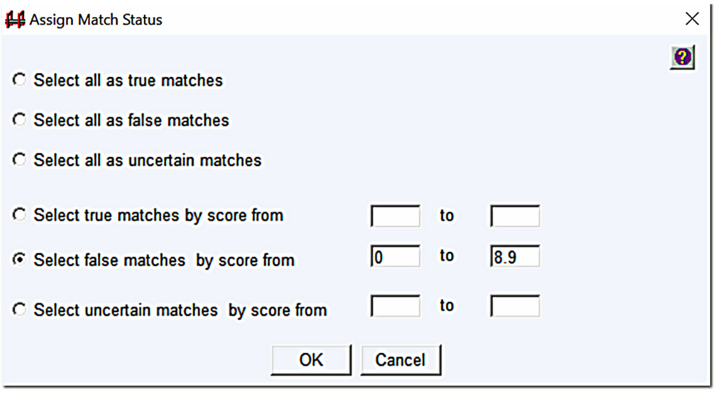
☑ = true mat

	Score	Class	Link ID	File	Reco
		1	1	1	1262
☑	40.7	1	1	2	1815
		1	2	1	5862
☑	40.7	1	2	2	5330
		1	3	1	2053
☑	40.7	1	3	2	6149
		1	4	1	2145
☑	40.7	1	4	2	6093
		1	5	1	1204

Click on the **Data** menu again, and select **Assign Match Status**… option a second time.

Result: The Assign Match Status dialogue box opens (Fig. WC.A1.35). Click on **Select false matches by score from**, enter **0** in the first box, enter **8.9** in the second box, and then click on **OK**.

Fig. WC.A1.35 Assigning cutoff of false matches



Assign Match Status

○ Select all as true matches

○ Select all as false matches

○ Select all as uncertain matches

○ Select true matches by score from [ ] to [ ]

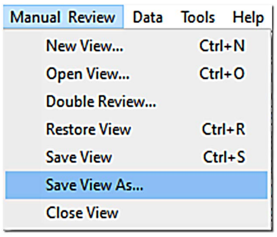
● Select false matches by score from [0] to [8.9]

○ Select uncertain matches by score from [ ] to [ ]

Any review session may be saved as a .view file and reopened at a later time. The steps for saving a review session are outlined below.

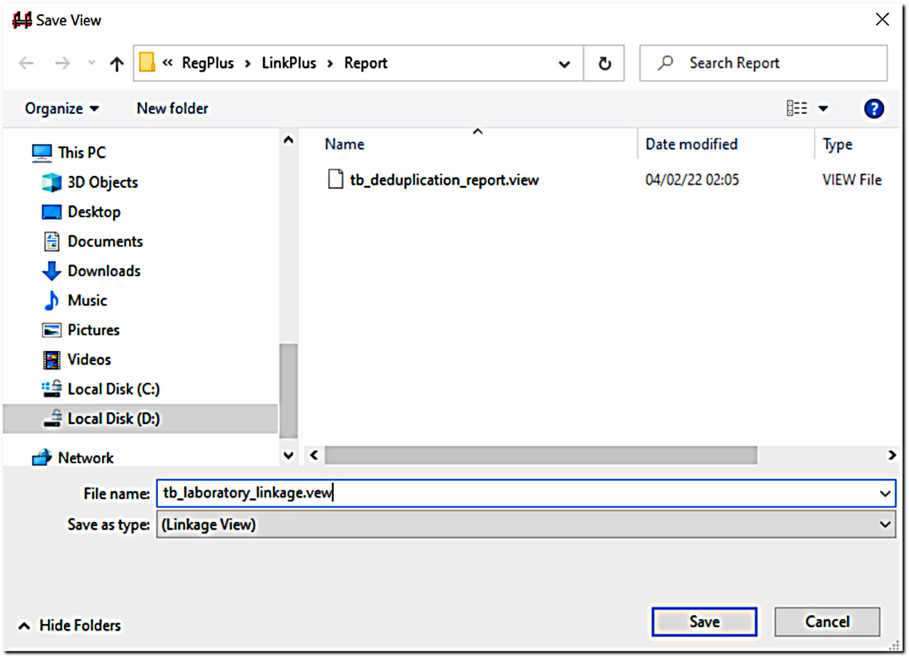
Click on the **Manual Review** menu and select **Save View As**… (Fig. WC.A1.36).

Fig. WC.A1.36 Saving review session as a .view file



Result: The **Save View** dialogue box opens (Fig. WC.A1.37).

Fig. WC.A1.37 Save View dialogue box



Save View

« RegPlus › LinkPlus › Report

Name	Date modified	Type
tb_deduplication_report.view	04/02/22 02:05	VIEW File

File name: tb_laboratory_linkage.vew

Save as type: (Linkage View)

Enter **tb_laboratory_linkage.view** in the File name box.

Click on **Save**.

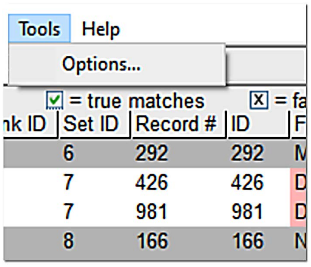
To reopen a saved review session, complete these steps: Click on the **Manual Review** menu and select **Open View**, then right-click on the **tb_laboratory_linkage.view** file and click on **Open**.

Step 11: Export merged file

Link Plus allows users to export the linkage file in a text file in delimited file format, specify the file delimiter and specify the match status(es) for export. The steps for exporting the deduplication are outlined below.

Click on the **Tools** menu, and select **Options…** (Fig. WC.A1.38).

Fig. WC.A1.38 Opening the Options dialogue box

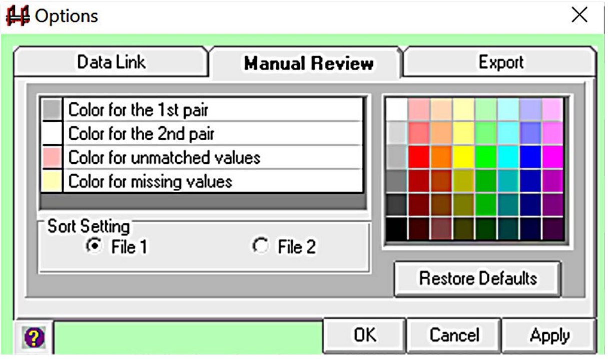


☑ = true matches ☒ = fa

nk ID	Set ID	Record #	ID	F
	6	292	292	M
	7	426	426	D
	7	981	981	D
	8	166	166	N

Result: The **Options** dialogue box opens, with the **Manual Review** tab selected (Fig. WC.A1.39).

Fig. WC.A1.39 Options dialogue box (Manual Review tab)



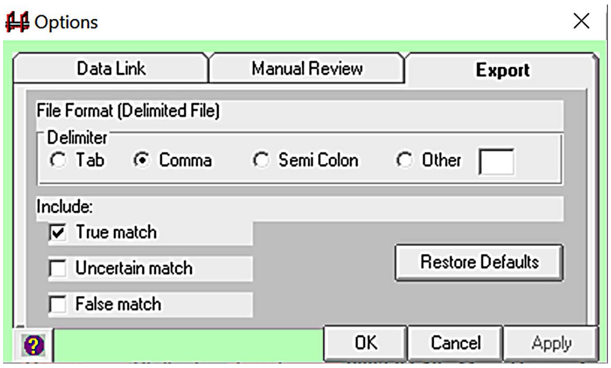
Color for the 1st pair
Color for the 2nd pair
Color for unmatched values
Color for missing values

Sort Setting
● File 1 ○ File 2

Click on the **Export** tab.

Result: The Export Options dialogue box is prompted (Fig. WC.A1.40).

Fig. WC.A1.40 Export Options dialogue box



File Format (Delimited File)

Delimiter
○ Tab ● Comma ○ Semi Colon ○ Other

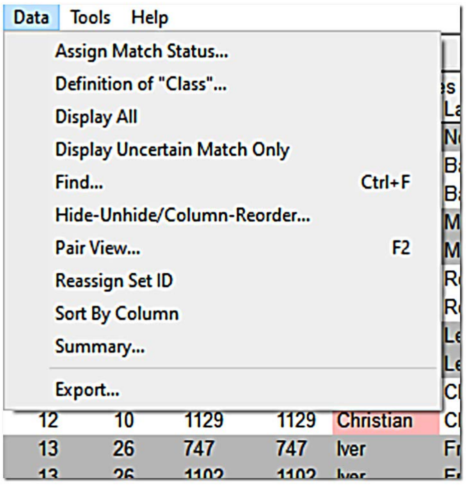
Include:
☑ True match
□ Uncertain match
□ False match

To set the File Format option to comma-delimited, click next to **Comma**. Unclick the **Uncertain match** and **False match** tick boxes, then click on **OK**.

The export file format has now been set to a comma-delimited file format, and when the data are exported, only duplicate records (i.e. those assigned a match status of true) will be exported.

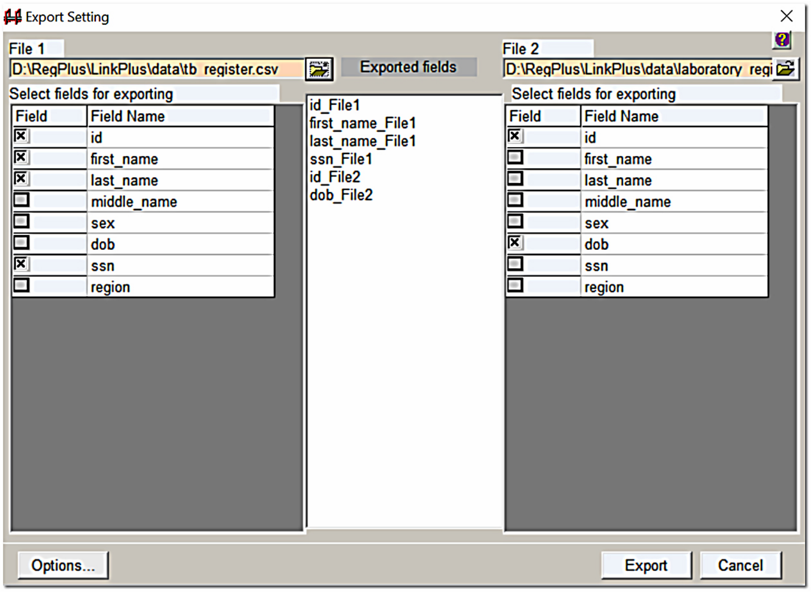
To export the file, click on the **Data** menu, and select **Export**… (Fig. WC.A1.41).

Fig. WC.A1.41 Data menu



Result: The Export Setting dialogue box opens (Fig. WC.A1.42).

Fig. WC.A1.42 Export Setting dialogue box



File 1
D:\RegPlus\LinkPlus\data\tb_register.csv

Select fields for exporting

Field	Field Name
☒	id
☒	first_name
☒	last_name
□	middle_name
□	sex
□	dob
☒	ssn
□	region

Exported fields:
id_File1
first_name_File1
last_name_File1
ssn_File1
id_File2
dob_File2

File 2
D:\RegPlus\LinkPlus\data\laboratory_regi

Select fields for exporting

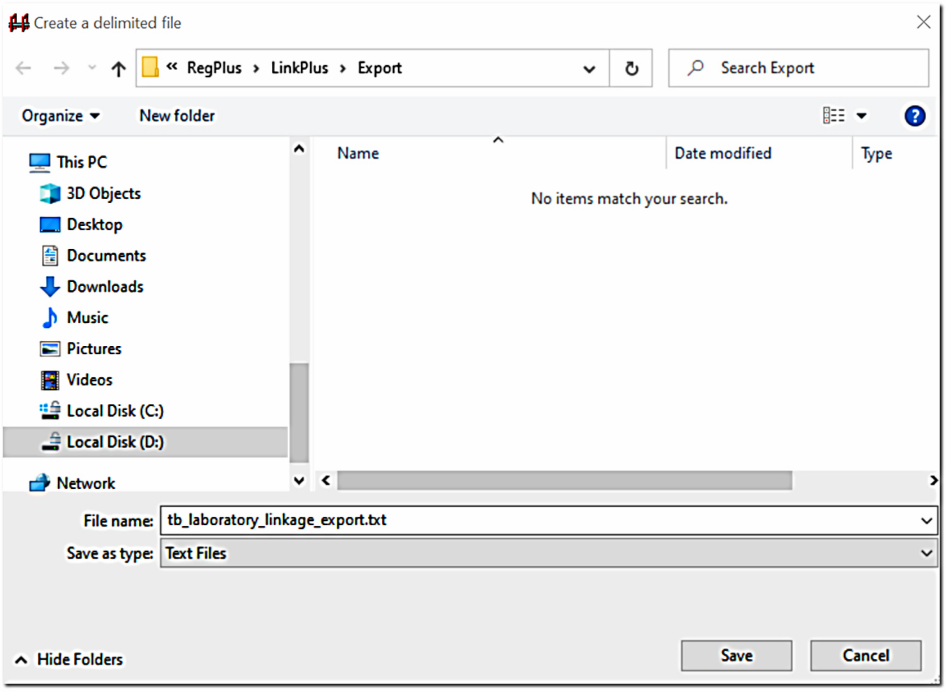
Field	Field Name
☒	id
□	first_name
□	last_name
□	middle_name
□	sex
☒	dob
□	ssn
□	region

Select the fields from File 1 and File 2 that you would like in your merged export file by ticking the boxes in the column to the left of the data field. Select the fields in the order in which you would like them listed in the export file.

Click on the **Export** button.

Result: A window will open where you can specify a name for the exported merged file, and a location on your PC or on a shared network folder where you would like to save the file. The default location for exported files within Link Plus is the C:/RegPlus/LinkPlus/Export folder (Fig. WC.A1.43).

Fig. WC.A1.43 Saving the merged file in the export folder



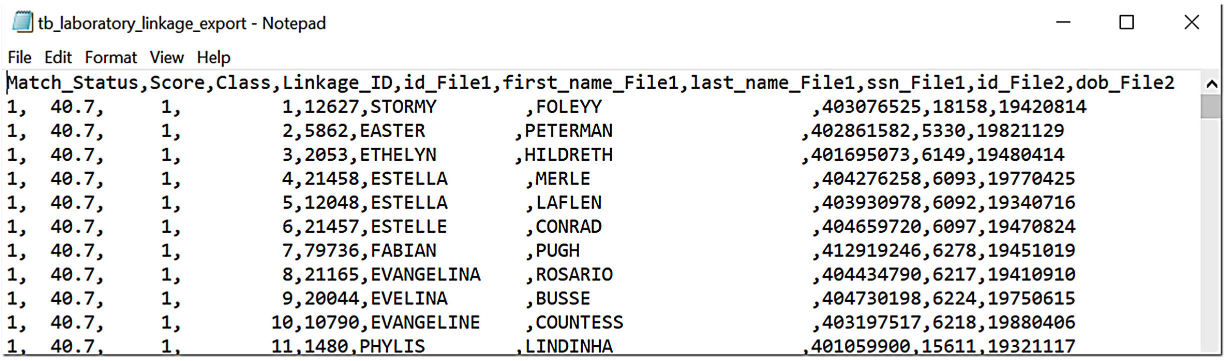
File name: tb_laboratory_linkage_export.txt

Save as type: Text Files

In the File name box, type **tb_laboratory_linkage_export.txt**, and click on the **Save** button.

Result: A comma-delimited merged file (named **tb_laboratory_linkage_export.txt**) of records with true match status, with the data fields you selected (in the order you selected them) is generated and placed in the C:/RegPlus/LinkPlus/Export folder. In addition, Notepad will open with a view of the merged file (Fig. WC.A1.44).

Fig. WC.A1.44 Saved review session as a .view file open in Notepad



tb_laboratory_linkage_export - Notepad
File Edit Format View Help

```
Match_Status,Score,Class,Linkage_ID,id_File1,first_name_File1,last_name_File1,ssn_File1,id_File2,dob_File2
1,  40.7,     1,          1,12627,STORMY        ,FOLEYY                   ,403076525,18158,19420814
1,  40.7,     1,          2,5862,EASTER        ,PETERMAN                 ,402861582,5330,19821129
1,  40.7,     1,          3,2053,ETHELYN       ,HILDRETH                 ,401695073,6149,19480414
1,  40.7,     1,          4,21458,ESTELLA       ,MERLE                    ,404276258,6093,19770425
1,  40.7,     1,          5,12048,ESTELLA       ,LAFLEN                   ,403930978,6092,19340716
1,  40.7,     1,          6,21457,ESTELLE       ,CONRAD                   ,404659720,6097,19470824
1,  40.7,     1,          7,79736,FABIAN        ,PUGH                     ,412919246,6278,19451019
1,  40.7,     1,          8,21165,EVANGELINA    ,ROSARIO                  ,404434790,6217,19410910
1,  40.7,     1,          9,20044,EVELINA       ,BUSSE                    ,404730198,6224,19750615
1,  40.7,     1,         10,10790,EVANGELINE    ,COUNTESS                 ,403197517,6218,19880406
1,  40.7,     1,         11,1480,PHYLIS        ,LINDINHA                 ,401059900,15611,19321117
```

Close Notepad and click on **Cancel** in the Export Setting dialogue box to return to the manual review screen. Click on the **Manual Review** menu and select the **Close View** option.

Annex 2

Instructions for deduplication using the Link Plus application

Adapted from Link Plus Stand-Alone Probabilistic Record Linkage Software, Linkage exercises, Oregon, May 2006

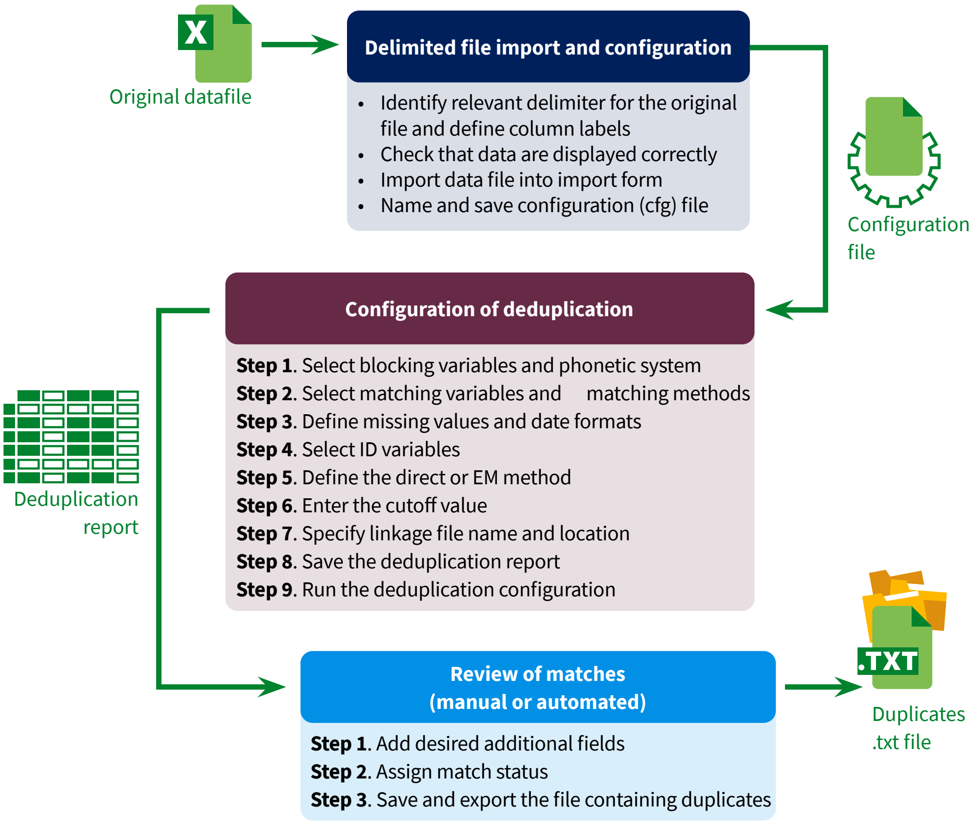
Link Plus is a probabilistic record-linkage program developed at the Cancer Division of the United States Centers for Disease Control and Prevention (US CDC). The main advantage of Link Plus is that it allows partial (approximate) matches on character variables (patient names) and fuzzy matching using phonetic coding systems as well as other variable-specific matching methods.

Getting started

- Download and install Link Plus software (www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm).
- The file to be investigated for probable duplicate records (**tb_ deduplication**) can be downloaded from here (link). Save these files in the subfolder RegPlus > LinkPlus > data folder as .csv files (comma-delimited).
- Open Local Disk (D:)> RegPlus > LinkPlus > data.

The deduplication process has three stages, each of which comprises several steps and involves input and output products. The output of one stage is used as input file for the following stage of the deduplication procedure (Fig. WC.A2.1).

Fig. WC.A2.1 Overview of the deduplication process using Link Plus



Delimited file import and configuration

- Identify relevant delimiter for the original file and define column labels
- Check that data are displayed correctly
- Import data file into import form
- Name and save configuration (cfg) file

Configuration of deduplication

Step 1. Select blocking variables and phonetic system
Step 2. Select matching variables and matching methods
Step 3. Define missing values and date formats
Step 4. Select ID variables
Step 5. Define the direct or EM method
Step 6. Enter the cutoff value
Step 7. Specify linkage file name and location
Step 8. Save the deduplication report
Step 9. Run the deduplication configuration

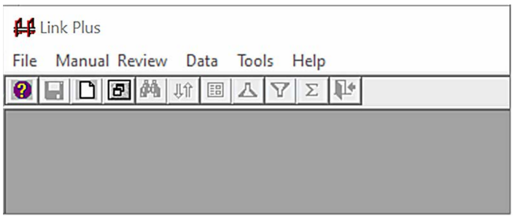
Review of matches (manual or automated)

Step 1. Add desired additional fields
Step 2. Assign match status
Step 3. Save and export the file containing duplicates

Launching Link Plus and starting the deduplication exercise

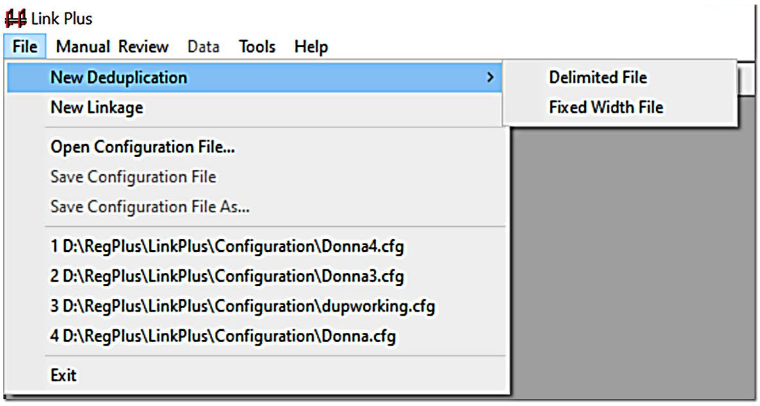
Launch the Link Plus application by clicking on Start>All Programs>Registry Plus>Link Plus>Link Plus. The main Link Plus window opens (Fig. WC.A2.2).

Fig. WC.A2.2 Link Plus window



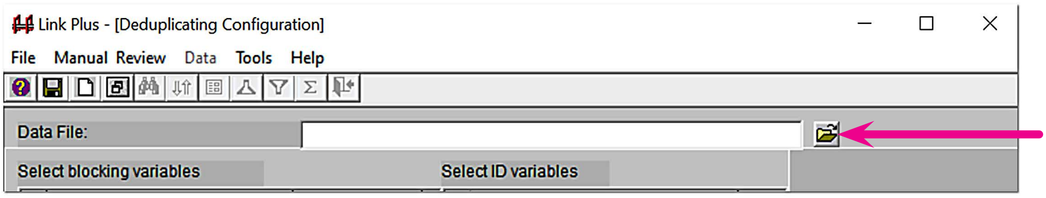
Go to the **File** menu and select **New Deduplication**, then **Delimited File** (Fig. WC.A2.3).

Fig. WC.A2.3 Link Plus menu


Link Plus
File
Manual Review
Data
Tools
Help
New Deduplication
New Linkage
Open Configuration File...
Save Configuration File
Save Configuration File As...
1 D:\RegPlus\LinkPlus\Configuration\Donna4.cfg
2 D:\RegPlus\LinkPlus\Configuration\Donna3.cfg
3 D:\RegPlus\LinkPlus\Configuration\dupworking.cfg
4 D:\RegPlus\LinkPlus\Configuration\Donna.cfg
Exit
Delimited File
Fixed Width File

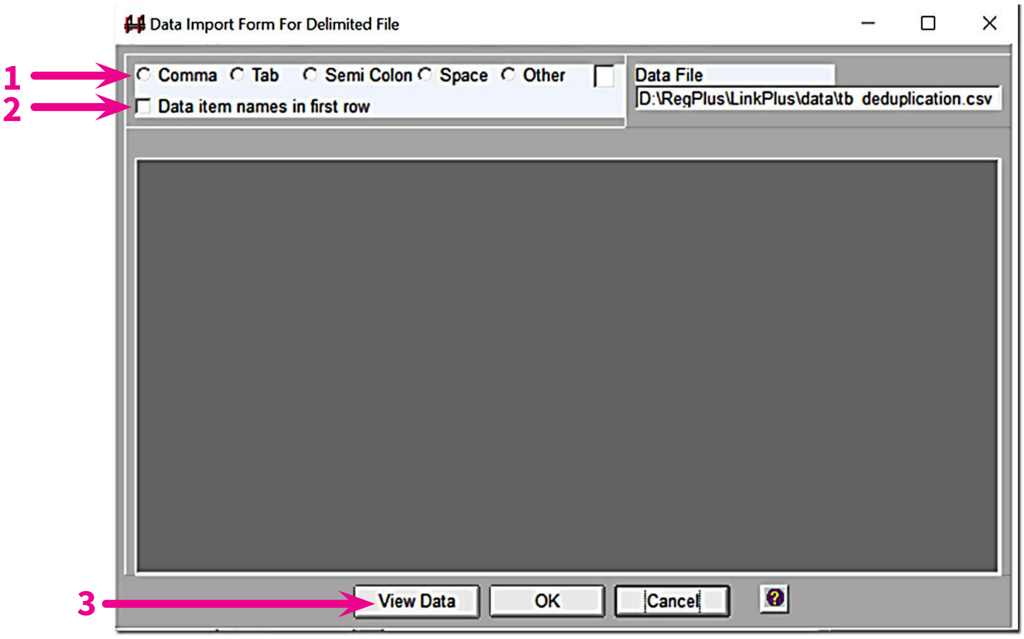
Click on the **folder** icon to the right of the Data File input box (Fig. WC.A2.4).

Fig. WC.A2.4 Data file input box


Link Plus - [Deduplicating Configuration]
File
Manual Review
Data
Tools
Help
Data File:
Select blocking variables
Select ID variables

Select the **tb_deduplication.csv** file and click on the **Open** button. The **Data Import Form For Delimited File** window will open. In our example, **tb_deduplication.csv** is a comma-delimited file and the first row contains the variable names; therefore, tick the **Comma** option button (number 1 in Fig. WC.A2.5) and **Data item names in first row** box (number 2 in Fig. WC.A2.5) and then click on **View Data** (number 3 in Fig. WC.A2.5).

Fig. WC.A2.5 Data Import Form for Delimited File



Data Import Form For Delimited File

○ Comma ○ Tab ○ Semi Colon ○ Space ○ Other □

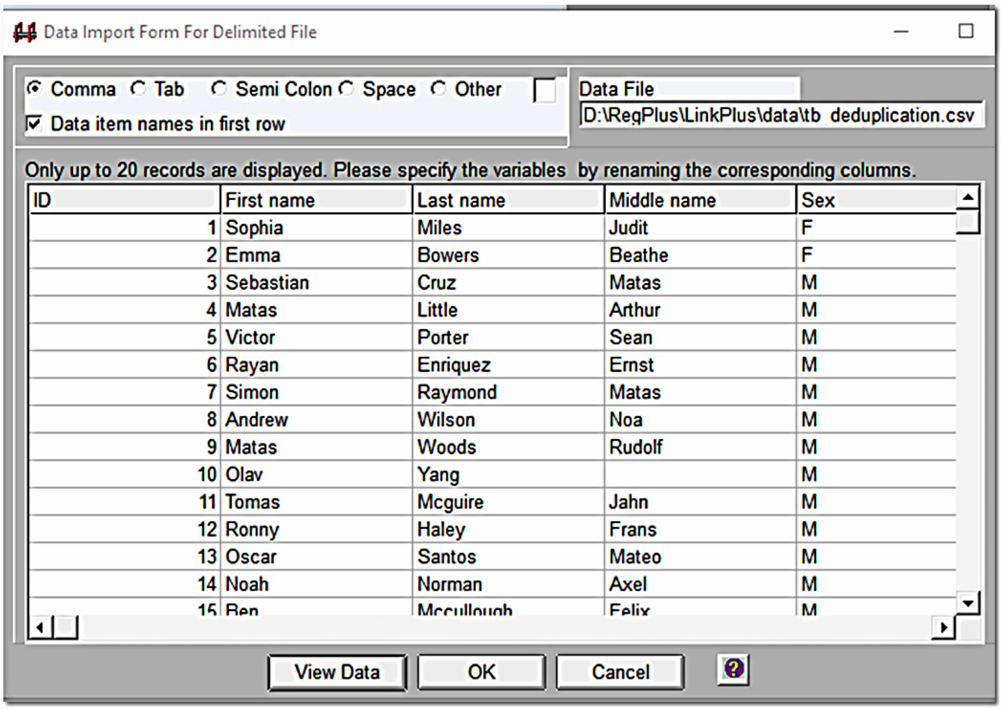
□ Data item names in first row

Data File

D:\RegPlus\LinkPlus\data\tb deduplication.csv

Check whether the data are displayed correctly in their proper columns (Fig. WC.A2.6).

Fig. WC.A2.6 Data are imported into import form and all columns are properly displayed



Data Import Form For Delimited File

● Comma ○ Tab ○ Semi Colon ○ Space ○ Other □

☑ Data item names in first row

Data File

D:\RegPlus\LinkPlus\data\tb deduplication.csv

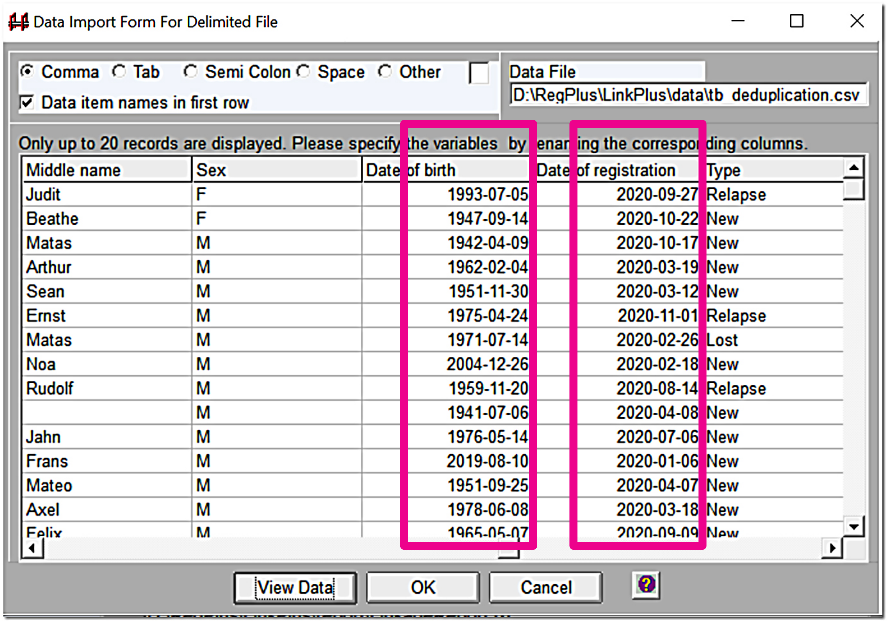
Only up to 20 records are displayed. Please specify the variables by renaming the corresponding columns.

ID	First name	Last name	Middle name	Sex
1	Sophia	Miles	Judit	F
2	Emma	Bowers	Beathe	F
3	Sebastian	Cruz	Matas	M
4	Matas	Little	Arthur	M
5	Victor	Porter	Sean	M
6	Rayan	Enriquez	Ernst	M
7	Simon	Raymond	Matas	M
8	Andrew	Wilson	Noa	M
9	Matas	Woods	Rudolf	M
10	Olav	Yang		M
11	Tomas	Mcguire	Jahn	M
12	Ronny	Haley	Frans	M
13	Oscar	Santos	Mateo	M
14	Noah	Norman	Axel	M
15	Ben	Mccullough	Felix	M

Using the horizontal scroll bar, navigate to the right to make sure that all fields have been imported. If some appear jumbled or concatenated, then the delimiter option may be inappropriately selected and should be re-specified.

Note: Date formats may vary, depending on which spreadsheet was used to create the file. Check that the imported date fields are specified either in YYYYMMDD or MMDDYYYY format (Fig. WC.A2.7) – those are the only date formats allowed by Link Plus.

Fig. WC.A2.7 Data import form for delimited file showing date formats for checking



● Comma ○ Tab ○ Semi Colon ○ Space ○ Other □

☑ Data item names in first row

Data File: D:\RegPlus\LinkPlus\data\tb deduplication.csv

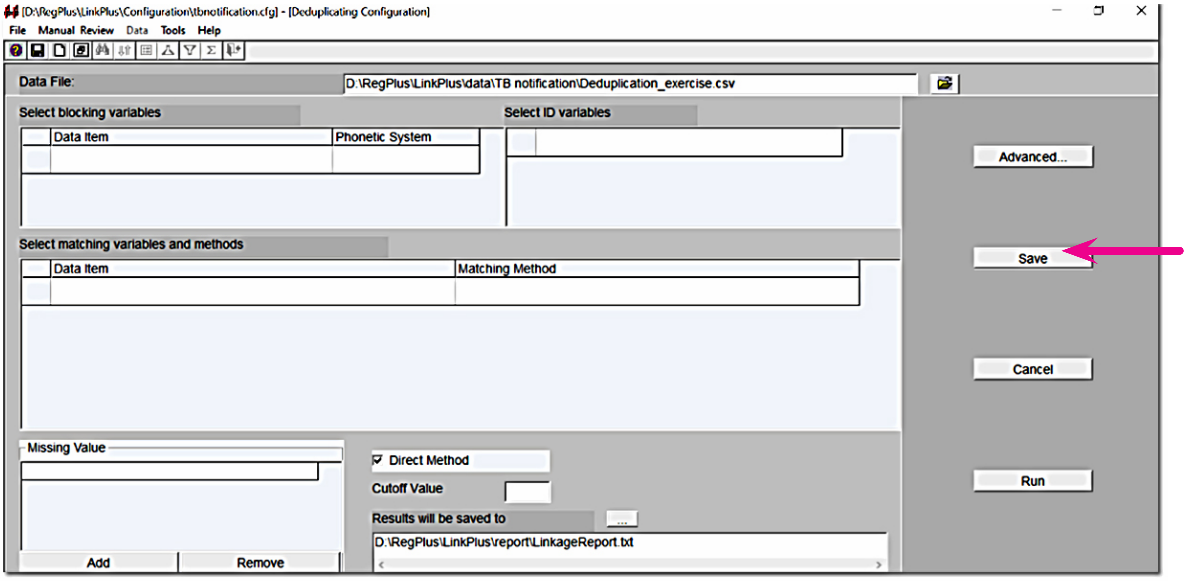
Only up to 20 records are displayed. Please specify the variables by renaming the corresponding columns.

Middle name	Sex	Date of birth	Date of registration	Type
Judit	F	1993-07-05	2020-09-27	Relapse
Beathe	F	1947-09-14	2020-10-22	New
Matas	M	1942-04-09	2020-10-17	New
Arthur	M	1962-02-04	2020-03-19	New
Sean	M	1951-11-30	2020-03-12	New
Ernst	M	1975-04-24	2020-11-01	Relapse
Matas	M	1971-07-14	2020-02-26	Lost
Noa	M	2004-12-26	2020-02-18	New
Rudolf	M	1959-11-20	2020-08-14	Relapse
	M	1941-07-06	2020-04-08	New
Jahn	M	1976-05-14	2020-07-06	New
Frans	M	2019-08-10	2020-01-06	New
Mateo	M	1951-09-25	2020-04-07	New
Axel	M	1978-06-08	2020-03-18	New
Felix	M	1965-05-07	2020-09-09	New

If everything looks fine in the Data Import Form For Delimited File, click on **OK.**

The Deduplicating Configuration window will be opened. Click on **Save** (Fig. WC.A2.8)**.**

Fig. WC.A2.8 The Deduplicating Configuration window


[D:\RegPlus\LinkPlus\Configuration\tbnotification.cfg] - [Deduplicating Configuration]
File
Manual Review
Data
Tools
Help
Advanced...
Save
Cancel
Run
Add
Remove

Data File: D:\RegPlus\LinkPlus\data\TB notification\Deduplication_exercise.csv

Select blocking variables

Data Item	Phonetic System

Select ID variables

Select matching variables and methods

Data Item	Matching Method

Missing Value

☑ Direct Method

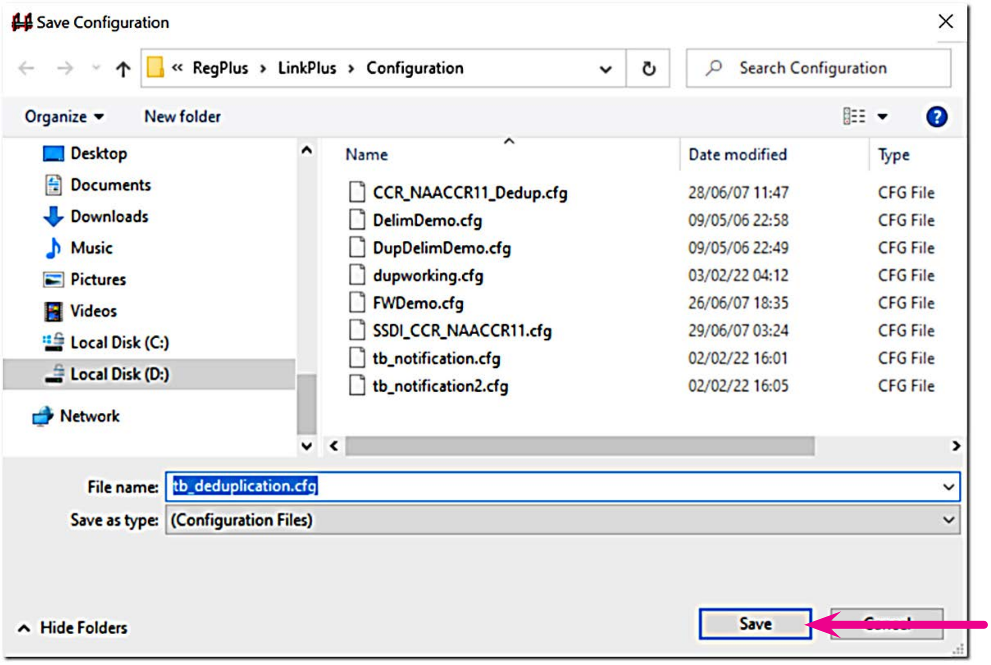
Cutoff Value

Results will be saved to

D:\RegPlus\LinkPlus\report\LinkageReport.txt

You will be prompted to save the configuration file in the default Configuration folder. Name the file **tb_deduplication.cfg** and click on **Save** (Fig. WC.A2.9).

Fig. WC.A2.9 Saving the configuration file in the Configuration folder



Save Configuration

« RegPlus › LinkPlus › Configuration

Search Configuration

Organize ▾   New folder

- Desktop
- Documents
- Downloads
- Music
- Pictures
- Videos
- Local Disk (C:)
- Local Disk (D:)
- Network

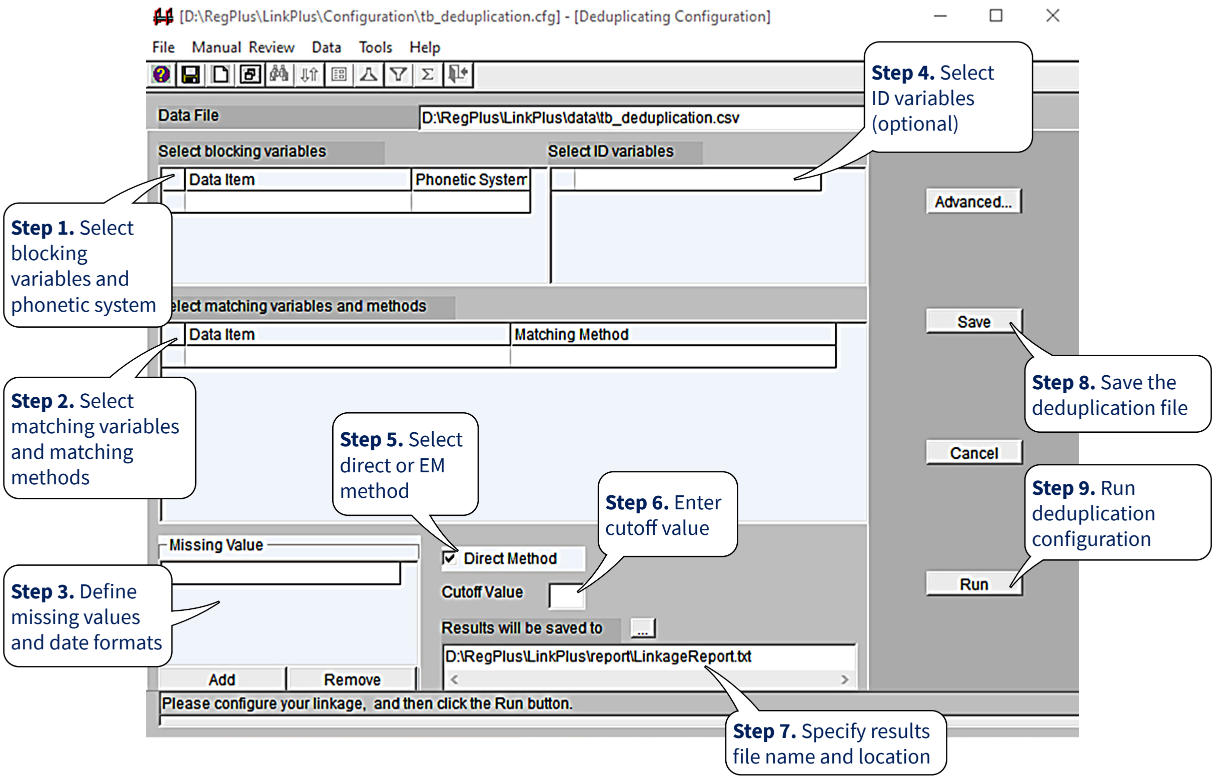
Name	Date modified	Type
CCR_NAACCR11_Dedup.cfg	28/06/07 11:47	CFG File
DelimDemo.cfg	09/05/06 22:58	CFG File
DupDelimDemo.cfg	09/05/06 22:49	CFG File
dupworking.cfg	03/02/22 04:12	CFG File
FWDemo.cfg	26/06/07 18:35	CFG File
SSDI_CCR_NAACCR11.cfg	29/06/07 03:24	CFG File
tb_notification.cfg	02/02/22 16:01	CFG File
tb_notification2.cfg	02/02/22 16:05	CFG File

File name: tb_deduplication.cfg

Save as type: (Configuration Files)

You are now ready to assess the file for deduplication following the steps shown in Fig. WC.A2.10.

Fig. WC.A2.10 Deduplicating Configuration dialogue box



[D:\RegPlus\LinkPlus\Configuration\tb_deduplication.cfg] - [Deduplicating Configuration]

Data File: D:\RegPlus\LinkPlus\data\tb_deduplication.csv

Select blocking variables

Data Item	Phonetic System

Select ID variables

Select matching variables and methods

Data Item	Matching Method

Missing Value

☑ Direct Method

Cutoff Value

Results will be saved to: D:\RegPlus\LinkPlus\report\LinkageReport.txt

Please configure your linkage, and then click the Run button.

Step 1. Select blocking variables and phonetic system

Step 2. Select matching variables and matching methods

Step 3. Define missing values and date formats

Step 4. Select ID variables (optional)

Step 5. Select direct or EM method

Step 6. Enter cutoff value

Step 7. Specify results file name and location

Step 8. Save the deduplication file

Step 9. Run deduplication configuration

Step 1: Select blocking variables and phonetic system

Blocking variables are the fields that are used in Link Plus to compare the pairs with identical values on at least one of those variables as a possible match, so they can be sent for comparison on matching variables. Common blocking variables include last name, first name, date of birth and national ID number. If any one of the blocking variables matches identically in the file, Link Plus will go on to compare match records and assign a score. You can select up to **five** fields for blocking.

Phonetic system: Phonetic coding involves coding a string on the basis of how it is pronounced. It allows for fuzzy matching of names based on what people at registration desks may hear, interpret and type in when registering a client. There are two phonetic coding systems in Link Plus:

- Soundex, which reduces matching problems due to different spellings and is simple and fast; it has fewer distinctive groupings than NYSIIS; and
- NYSIIS, which offers an improvement on the Soundex algorithm; for example, some studies suggest that NYSIIS performs better than Soundex when Spanish names are used.

Using the drop-down menus for the Data Item and Phonetic System columns, select the following blocking variables and phonetic systems for the **tb_deduplication.cfg** deduplication.

Data Item	Phonetic System
First name	Soundex
Last name	Soundex
Date of birth	

Step 2: Select matching variables and matching methods

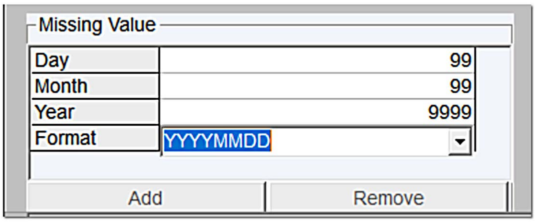
Choose the desired variables from the drop-down list. The selected row will be indicated by an asterisk on the left. You can select up to 10 fields for matching. Select the following matching variables and matching methods for the **tb-deduplication.cfg** deduplication exercise:

Data Item	Matching Method
First name	First name
Last name	Last name
Middle name	Middle name
Date of birth	Date
Sex	Exact
Type	Exact

Step 3: Define missing values and date formats

Link Plus automatically treats null or empty values as missing data for matching variables, and allows the user to indicate additional values that are to be treated as missing data by the program. Select the YYYYDDMM format for date of birth variables. The missing values specified for the **tb_deduplication.cfg** configuration are shown in Fig. WC.A2.11.

Fig. WC.A2.11 Defining date formats and missing values



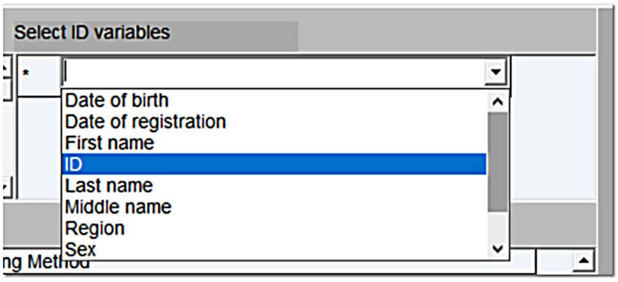
Missing Value

Day	99
Month	99
Year	9999
Format	YYYYMMDD

Step 4: Select ID variables (optional)

You may wish to define variables for identifying records to include in the reports. Selection of ID variables is optional. To select ID variables for the data file, click in the Select ID variables grid, and select the desired variable or variables from the drop-down list that appears (Fig. WC.A2.12).

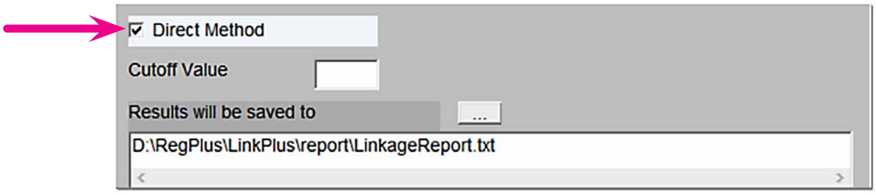
Fig. WC.A2.12 Selection of ID variables



Step 5: Select the direct or EM method

The M-probability is the probability that a matching variable agrees, given that a comparison pair is a match. For the **tb_deduplication** exercise, select Direct Method to derive the M-probabilities used in the linkage. When the Direct Method box is ticked, default M-probabilities are used in the linkage. Leave the Direct Method box ticked (Fig. WC.A2.13).

Fig. WC.A2.13 Selection of method to derive M-probability



☑ Direct Method

Cutoff Value

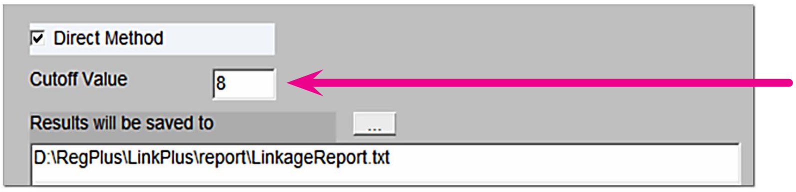
Results will be saved to ...

D:\RegPlus\LinkPlus\report\LinkageReport.txt

Step 6: Enter the cutoff value

The cutoff value is the score above which comparison pairs are accepted as potential duplicates and presented for manual review. Specify 8 as the cutoff value for this exercise. It is recommended that you set the cutoff value anywhere from 7 to 10 (Fig. WC.A2.14).

Fig. WC.A2.14 Input box to enter cutoff value



☑ Direct Method

Cutoff Value 8

Results will be saved to ...

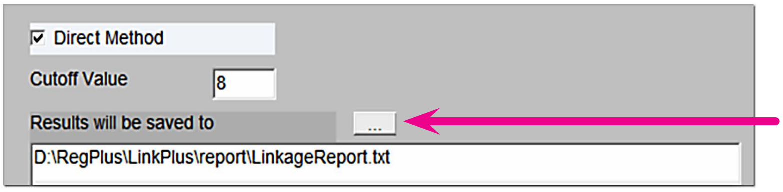
D:\RegPlus\LinkPlus\report\LinkageReport.txt

Step 7: Specify linkage file name and location

At the completion of a deduplication linkage run, Link Plus will generate a linkage report, named **LinkageReport.txt**, and store it in the report folder of the Link Plus directory. You will need to rename the linkage report, otherwise it will be overwritten the next time you run Link Plus. The steps used to rename the linkage report are set out below.

Click on the box to the right of **Results will be saved to** (Fig. WC.A2.15)**.**

Fig. WC.A2.15 Opening dialogue box for saving the deduplication report name and location



☑ Direct Method

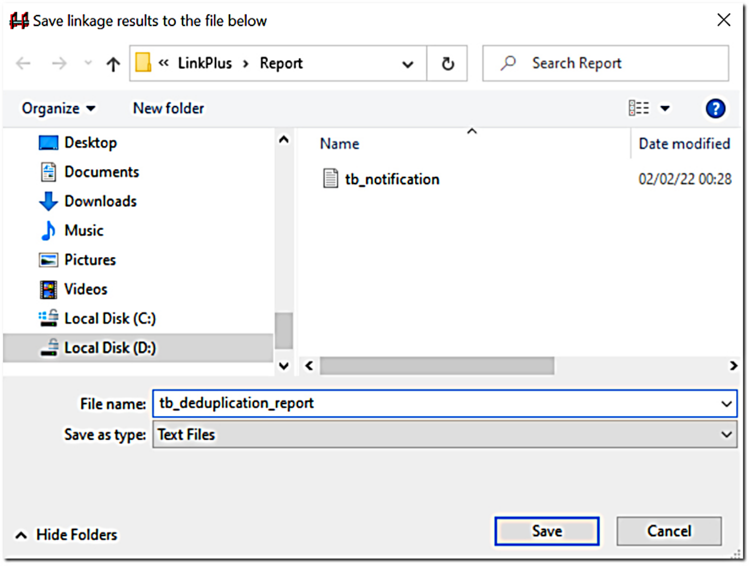
Cutoff Value: 8

Results will be saved to ...

D:\RegPlus\LinkPlus\report\LinkageReport.txt

The Save linkage results to the file below dialogue box is prompted (Fig. WC.A2.16).

Fig. WC.A2.16 Dialogue box for Save Linkage results to the file below



Save linkage results to the file below

« LinkPlus › Report

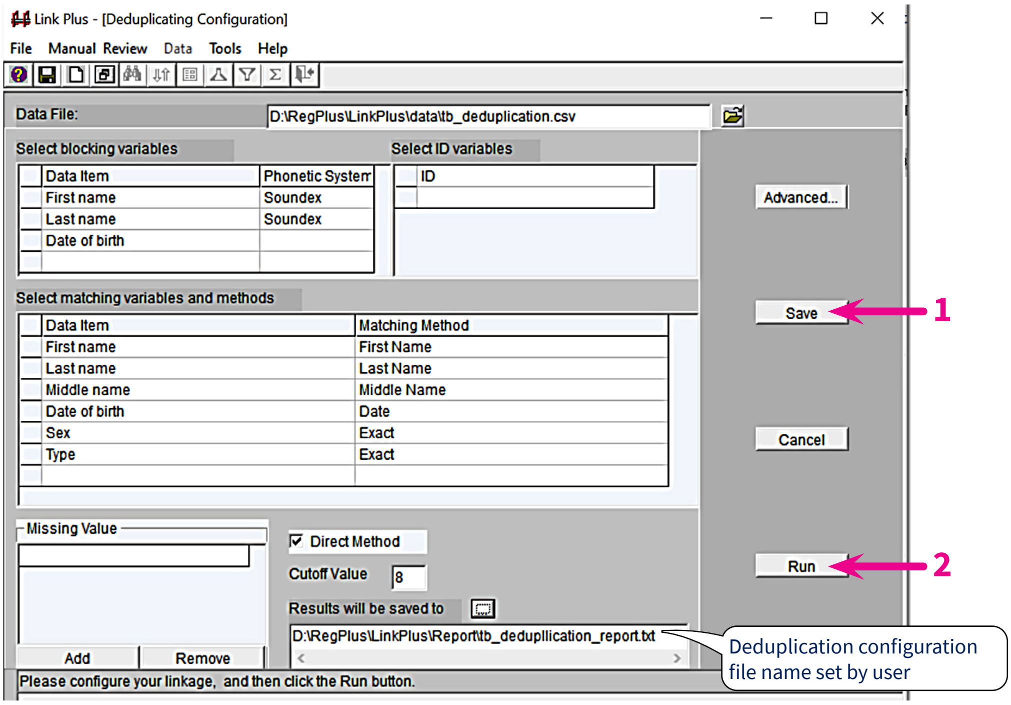
Name	Date modified
tb_notification	02/02/22 00:28

File name: tb_deduplication_report

Save as type: Text Files

Enter **tb_deduplication_report.txt** in the File name box, and click on **Save.** The application takes you back to the Deduplicating Configuration window, where the linkage report name has been changed to **tb_deduplication_report.txt** (Fig. WC.A2.17).

Fig. WC.A2.17 Saving the deduplication report and running the deduplication configuration



Link Plus - [Deduplicating Configuration]

Data File: D:\RegPlus\LinkPlus\data\tb_deduplication.csv

Select blocking variables

Data Item	Phonetic System
First name	Soundex
Last name	Soundex
Date of birth	

Select ID variables

ID

Select matching variables and methods

Data Item	Matching Method
First name	First Name
Last name	Last Name
Middle name	Middle Name
Date of birth	Date
Sex	Exact
Type	Exact

Missing Value

☑ Direct Method

Cutoff Value: 8

Results will be saved to

D:\RegPlus\LinkPlus\Report\tb_dedupllication_report.txt

Please configure your linkage, and then click the Run button.

Deduplication configuration file name set by user

Step 8: Save the deduplication report

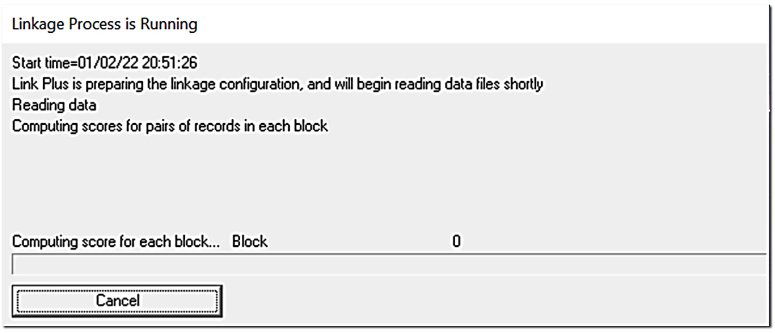
Click on **Save** to save the configuration file for this deduplication report.

Step 9: Run the deduplication configuration

Click on **Run** to run the linkage process.

The linkage process begins, and the Linkage Process progress window appears; this window provides the user with feedback about the linkage process as it is run (Fig. WC.A2.18).

Fig. WC.A2.18 Linkage Process progress window



Linkage Process is Running

Start time=01/02/22 20:51:26
Link Plus is preparing the linkage configuration, and will begin reading data files shortly
Reading data
Computing scores for pairs of records in each block

Computing score for each block... Block 0

The report is a tab-delimited text file, which can be opened with the manual review feature of Link Plus (it also can be opened in a text editor or spreadsheet program). Records are presented in comparison pairs, sorted by their linkage scores in descending order. Pairs with scores above the selected cutoff value are listed. The detail row for each record contains all of the matching and ID variables used in the linkage.

When the linkage is complete, the progress window will present you with the choice of conducting the manual review process directly after the linkage, or closing the progress window and conducting the manual review process at a later time.

A manual review of the identified duplicates can then be conducted as described in “Step 10: Manually review uncertain matches” in Annex 1.

Annex 3

Exercises in external linkage and deduplication, using the Link Plus application

In the following exercises, it is assumed that the preprocessing steps have already been done (e.g. cleaning and verification of the data in the input files, standardization of capital and lower case variables, and trimming of leading and training white spaces).

Exercise 1. Identifying laboratory monitoring records for TB patients after diagnosis

Objective: The objective of this exercise is to use Link Plus software to identify records in the laboratory register that represent the same patient in the TB register.

Launch the software using the steps in "Launching Link Plus and starting the external linkage exercise" in Annex 1, making sure to download the working files (.csv format) from link 1, link 2 and link 3. Check that you have selected the **Linkage** configuration. Now follow Steps 1 to 9 in Annex 1. The linkage process will take about 10 minutes to complete.

Question

Would you have obtained better matching if you had only specified date of birth as a blocking variable, rather than using the last name and first name? In what instances would this be useful?

Answer

If there's a high chance that individuals interchange their surnames and first names, then blocking using these fields would reduce the chances of matching.

Question

Would the region variable have been useful as a blocking variable?

Answer

In this simulation, the laboratory file does not have region data; therefore, using the region variable as a blocking variable would not have added to the efficiency of the matching process.

Once the linkage process is complete, follow Step 10 in Annex 1.

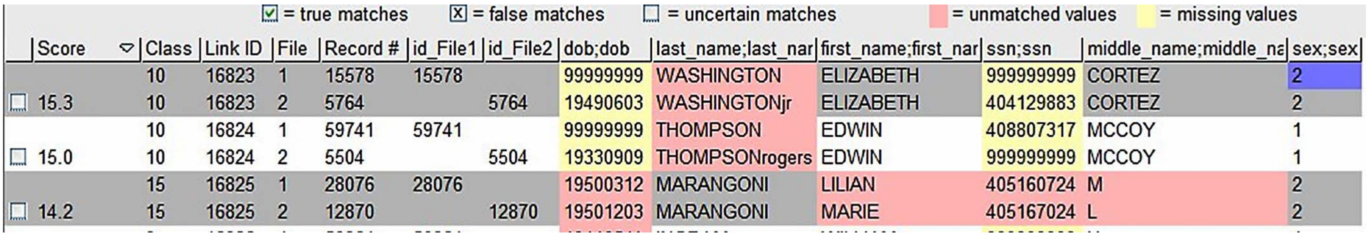
Question

The three records shown in Fig. WC.A3.1 have been matched. Looking at the record pair with a probability score of 14.2 (the third pair), what do you think is the reason for the low probability score compared with the two pairs above? Is this supposed to be a true match?

Answer

In this example, the middle and first names in the third pair of records are interchanged; also, the month and date of birth are interposed. This may be due to the use of different regional settings when creating the .csv files. Because four variables are highlighted as unmatched, the automatically computed score is lower than it is for the above pairs. Using an EM approach rather than a direct approach to compute the matching probability score may result in a higher match score.

Fig. WC.A3.1 Example records



☑ = true matches ☒ = false matches □ = uncertain matches 🟥 = unmatched values 🟨 = missing values

	Score	Class	Link ID	File	Record #	id_File1	id_File2	dob;dob	last_name;last_nar	first_name;first_nar	ssn;ssn	middle_name;middle_na	sex;sex
		10	16823	1	15578	15578		99999999	WASHINGTON	ELIZABETH	999999999	CORTEZ	2
□	15.3	10	16823	2	5764		5764	19490603	WASHINGTONjr	ELIZABETH	404129883	CORTEZ	2
		10	16824	1	59741	59741		99999999	THOMPSON	EDWIN	408807317	MCCOY	1
□	15.0	10	16824	2	5504		5504	19330909	THOMPSONrogers	EDWIN	999999999	MCCOY	1
		15	16825	1	28076	28076		19500312	MARANGONI	LILIAN	405160724	M	2
□	14.2	15	16825	2	12870		12870	19501203	MARANGONI	MARIE	405167024	L	2

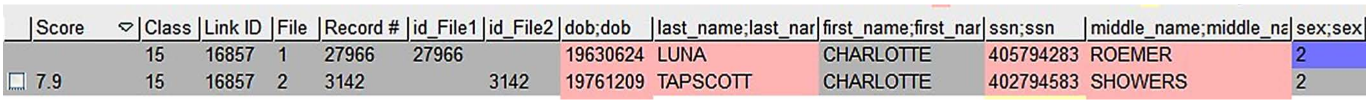
Question

Following a deduplication exercise, you obtain a result as shown in Fig. WC.A3.2 with a linkage probability score of 7.9. Would you consider this to be a false match or an uncertain match? Recalling that you set a matching threshold of 7, would you recommend increasing this threshold to 8?

Answer

In this example, only one of the five matching variables corresponds (i.e. the first name, "CHARLOTTE"). Hence, the records are very likely to be different. Since the social security numbers were very close (just by chance), the score is likely to be much higher than expected. Although it is possible to change the threshold, it may be helpful to start by manually reviewing the records that are near the threshold.

Fig. WC.A3.2 Sample records



	Score	Class	Link ID	File	Record #	id_File1	id_File2	dob;dob	last_name;last_nar	first_name;first_nar	ssn;ssn	middle_name;middle_na	sex;sex
		15	16857	1	27966	27966		19630624	LUNA	CHARLOTTE	405794283	ROEMER	2
□	7.9	15	16857	2	3142		3142	19761209	TAPSCOTT	CHARLOTTE	402794583	SHOWERS	2

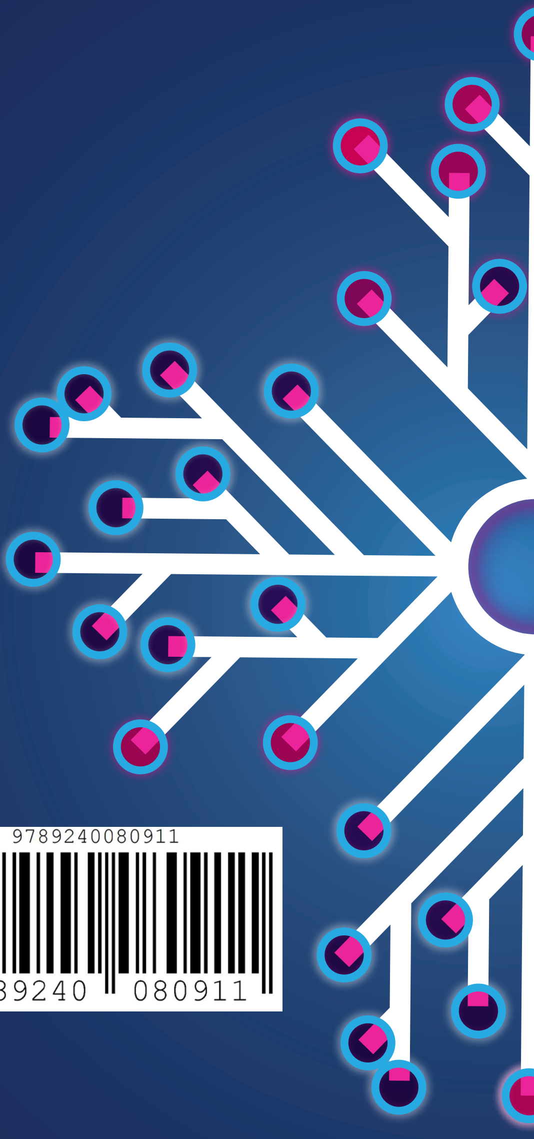
Exercise 2. Finding missed diagnoses

Using the manual review screen, open the file that contains the non-matched records from the laboratory. Display the record of the laboratory result.

Question

Which of these records would you consider important to follow up on, to determine whether they were missed diagnoses?

What would you consider to be the next step for the programme, to investigate these cases? What further variables would you need to assess from the laboratory register for these cases?



For further information, please contact:

Global Tuberculosis Programme

World Health Organization

20, Avenue Appia CH-1211 Geneva 27 Switzerland
Web site: www.who.int/tb

9789240080911

9 789240 080911