
Central Cancer Registry Perform Patient Linkage Use Case

Version 1.0

Prepared by: NPCR-AERRO Central Cancer Registry Workgroup
NPCR-AERRO Hospital Registry Workgroup
NPCR-AERRO Technical Development Team

Centers for Disease Control and Prevention
National Center for Chronic Disease Prevention and Health Promotion
Division of Cancer Prevention and Control
National Program of Cancer Registries

September 22, 2011



Table of Contents

General Information.....	3
Perform Patient Linkage	4
1.0 Precondition	4
2.0 Post Condition	4
3.0 Priority	4
4.0 Frequency of Use	4
5.0 Normal Course of Events.....	5
6.0 Alternative Course of Events.....	7
7.0 Business Rules.....	8
8.0 Exceptions.....	11
9.0 Includes.....	11
10.0 Special Requirements.....	11
11.0 Assumption.....	11
12.0 Notes and Issues	11
13.0 References	11
Appendix A: Perform Patient Linkage Workflow Diagram	12
Appendix B: Perform Patient Linkage Data Flow Diagram	13
Appendix C: References for Probabilistic Record Linkage	14
Appendix D: Matching Methods Used in Link Plus.....	15
Use Case Administrative Information	17

General Information

1. Use Case ID

CCRUC 1.4

2. Use Case Name

Perform Patient Linkage

3. Description

This use case describes the process of using defined criteria to determine whether source records refer to the same patient, based on the degree of agreement between demographic and other data fields. This process can be automated, manual, or a combination.

4. Actors

- Cancer registry (CR) software
- Registrar

5. Definitions

- **Event report:** An electronic transmission of information to a cancer registry.
- **Abstracted event report:** An extraction or summary of information created by a data source specifically for a cancer registry.
- **Electronic health record (EHR) event report:** A report, document, or note within the EHR, including radiology reports, pathology reports, clinician and nurses' notes, discharge summaries, and admissions forms.
- **Consolidated record:** The complete set of information on a patient, compiled from one or more event reports from one or more data sources.
- **Linkage score:** For a comparison pair, the overall weight of matching variables; a higher score means a greater likelihood of being a match.
- **Cutoff point (threshold):** The linkage score at which a comparison is considered a match or a non-match.
- **Match:** Records that are for the same person.
- **Non-match:** Records that are not for the same person.

Perform Patient Linkage

Note: A diagram for this use case is in [Appendix A](#).

1.0 Precondition

A set of conditions that must be met before the activities described in the use case can begin.

An event report has been validated and added to the central registry (CR) database a source record.

2.0 Post Condition

A set of conditions that must be met after the activities described in the use case have been completed.

The event report has been assigned a patient identification (ID) number.

3.0 Priority

Describes the importance and sequence of the use case in the overall activities of the cancer registry.

This is a high-priority use case.

4.0 Frequency of Use

Describes how often the activities in the use case take place.

The activities in this use case take place after an event report has been validated.

5.0 Normal Course of Events

Describes the specific steps taken to perform the activity in the use case.

Normal refers to the steps that are taken when everything goes according to routine procedures. Problems and exceptions are described in section 6, [Alternative Course](#).

Business rules are statements that describe a decision that must be made and agreed to by those involved in the activity. In the context of this document, a business rule describes the decision that needs to be made, and in some circumstances provides a recommendation; in others, options for consideration and use.

Software requirements are statements that describe the functionality of the software that is required or recommended.

5.1 The use case begins after the event report has been validated and added to the cancer registry (CR) database.

5.2 CR software compares the event report to records in the CR database and determines the linkage score. [BR01, BR02, BR03]

BR	Business Rule	Purpose	Remarks
01	The event report may be compared against the consolidated patient records or the existing event reports in the CR database.	To determine whether the patient already has a record in the registry.	
02	A probabilistic methodology algorithm should be used to determine the linkage score. Note: Establishing the weights, setting thresholds and calibrating probabilistic methodology is outside the scope of NPCR-AERRO.	To ensure accurate matching.	Link Plus is a freely available probabilistic patient linkage tool developed by CDC-NPCR. Refer to Appendix C for a list of references on probabilistic record linkage, and Appendix D for matching methods. <i>NAACCR Standards Volume III</i> also provides a discussion on patient linkage.
03	Data items for patient linkage include, at a minimum: <ul style="list-style-type: none"> • First name • Last name • Date of birth • Social Security number • Sex 	To ensure accurate matching.	Refer to <i>NAACCR Standards Volume III</i> for a discussion on key data items.

5.3 CR software assigns a patient identification (ID) number when the linkage score falls within the threshold for determining a match. [BR04, BR05, BR06, BR07, BR08]

5.3.1 CR software assigns the same patient ID number to the event report when the linkage score is within the threshold.

5.3.2 CR software assigns a new patient ID number to the event report when the linkage score is outside the threshold.

BR	Business Rule	Purpose	Remarks
04	The threshold should be set based on registrar review of a group of test cases.	To ensure accurate matching.	Methods for setting the thresholds used for probabilistic linkage are outside the scope of NPCR-AERRO and this use case.
05	Upper and lower thresholds should be set to minimize the number of false negative matches.	To ensure accurate matching.	Records are considered a match when their linkage scores fall between the upper and lower thresholds, and a non-match when their linkage scores fall outside the upper and lower thresholds. A false negative match occurs when the two records are determined to represent different people when they actually represent the same person.
06	Thresholds should be set based on— <ul style="list-style-type: none"> • Available data items (more data items allow you to set the thresholds more accurately, giving more confidence in the result). • Quality of data items (how often they are unknown and how accurate they are). 	To ensure accurate matching.	
07	Thresholds should be monitored and adjusted routinely to minimize false negative results.	To ensure accurate matching.	
08	The CCR should use the help files provided by their probabilistic linkage software to set the thresholds.	To ensure accurate matching.	

5.4 The registrar reviews the results of patient linkage for an event report when the score falls within the threshold for determining a match. [BR09, BR10, BR11]

BR	Business Rule	Purpose	Remarks
09	Additional data items that could be used to link the event report include— <ul style="list-style-type: none"> • Middle name • Maiden name • Alias • Race • Addresses (diagnosis and current) • Physician • Primary site 	To improve the accuracy of matching.	
10	The registrar may need to contact the data source or a prior event report to help link the new event report.	To ensure accurate matching.	
11	The following resources could be used to link the event report— <ul style="list-style-type: none"> • Internet phone directories • Social Security Death index • Birth and death certificates • Department of Motor Vehicles • Voter registration 	To ensure accurate matching.	

5.5 The registrar assigns the appropriate patient ID number.

5.6 CR software records the patient ID number in the source record.

5.7 The use case ends.

6.0 Alternative Course of Events

Not available.

7.0 Business Rules

A statement that describes a decision that must be made and agreed to by those involved in the activity. In the context of this document, a business rule describes the decision that needs to be made, and in some circumstances provides a recommendation; in others, options for consideration and use.

Business rules for this use case are presented under the step to which they apply.

BR	Business Rule	Purpose	Remarks
01	The event report may be compared against the consolidated patient records or the existing event reports in the CR database.	To determine whether the patient already has a record in the registry.	
02	A probabilistic methodology algorithm should be used to determine the linkage score. Note: Establishing the weights, setting thresholds and calibrating probabilistic methodology is outside the scope of NPCR-AERRO.	To ensure accurate matching.	Link Plus is a freely available probabilistic patient linkage tool developed by CDC-NPCR. Refer to Appendix C for a list of references on probabilistic record linkage, and Appendix D for matching methods. <i>NAACCR Standards Volume III</i> also provides a discussion on patient linkage.
03	Data items for patient linkage include, at a minimum: <ul style="list-style-type: none"> • First name • Last name • Date of birth • Social Security number • Sex 	To ensure accurate matching.	Refer to <i>NAACCR Standards Volume III</i> for a discussion on key data items.
04	The threshold should be set based on registrar review of a group of test cases.	To ensure accurate matching.	Methods for setting the thresholds used for probabilistic linkage are outside the scope of NPCR-AERRO and this use case.
05	Upper and lower thresholds should be set to minimize the number of false negative matches.	To ensure accurate matching.	Records are considered a match when their linkage scores fall between the upper and lower thresholds, and a non-match when their linkage scores fall outside the upper and lower thresholds. A false negative match occurs when the two records are determined to represent different people when they actually represent the same person.

BR	Business Rule	Purpose	Remarks
06	Thresholds should be set based on— <ul style="list-style-type: none"> • Available data items (more data items allow you to set the thresholds more accurately, giving more confidence in the result). • Quality of data items (how often they are unknown and how accurate they are). 	To ensure accurate matching.	
07	Thresholds should be monitored and adjusted routinely to minimize false negative results.	To ensure accurate matching.	
08	The CCR should use the help files provided by their probabilistic linkage software to set the thresholds.	To ensure accurate matching.	
09	Additional data items that could be used to link the event report include— <ul style="list-style-type: none"> • Middle name • Maiden name • Alias • Race • Addresses (diagnosis and current) • Physician • Primary site 	To improve the accuracy of matching.	
10	The registrar may need to contact the data source or a prior event report to help link the new event report.	To ensure accurate matching.	

BR	Business Rule	Purpose	Remarks
11	The following resources could be used to link the event report— <ul style="list-style-type: none">• Internet phone directories• Social Security Death index• Birth and death certificates• Department of Motor Vehicles• Voter registration	To ensure accurate matching.	

8.0 Exceptions

None.

9.0 Includes

None.

10.0 Special Requirements

None.

11.0 Assumption

Batch files are in an electronic format.

12.0 Notes and Issues

None.

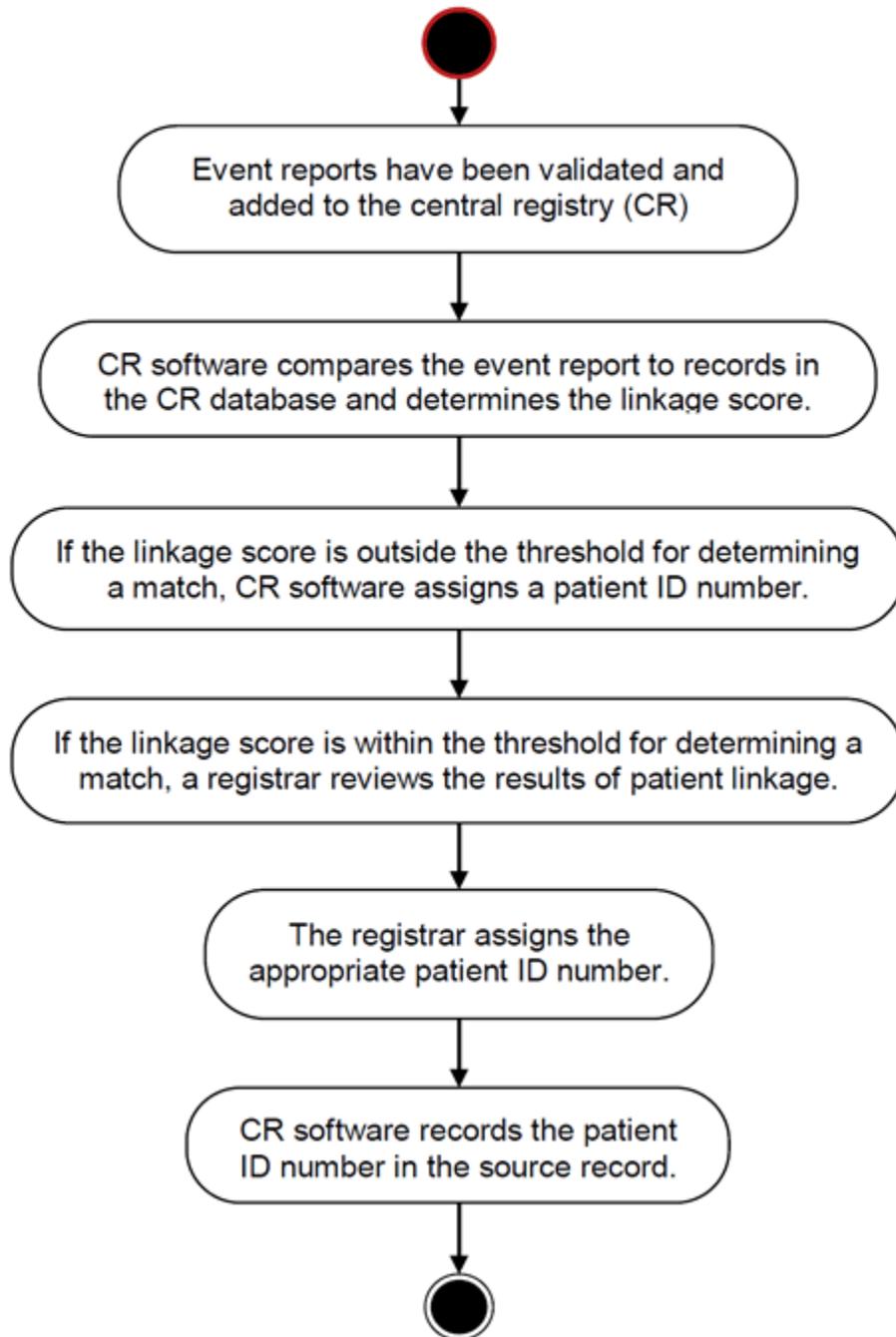
13.0 References

North American Association of Central Cancer Registries Standards for Cancer Registries, Volume III: Standards for Completeness, Quality, Analysis, and Management of Data.

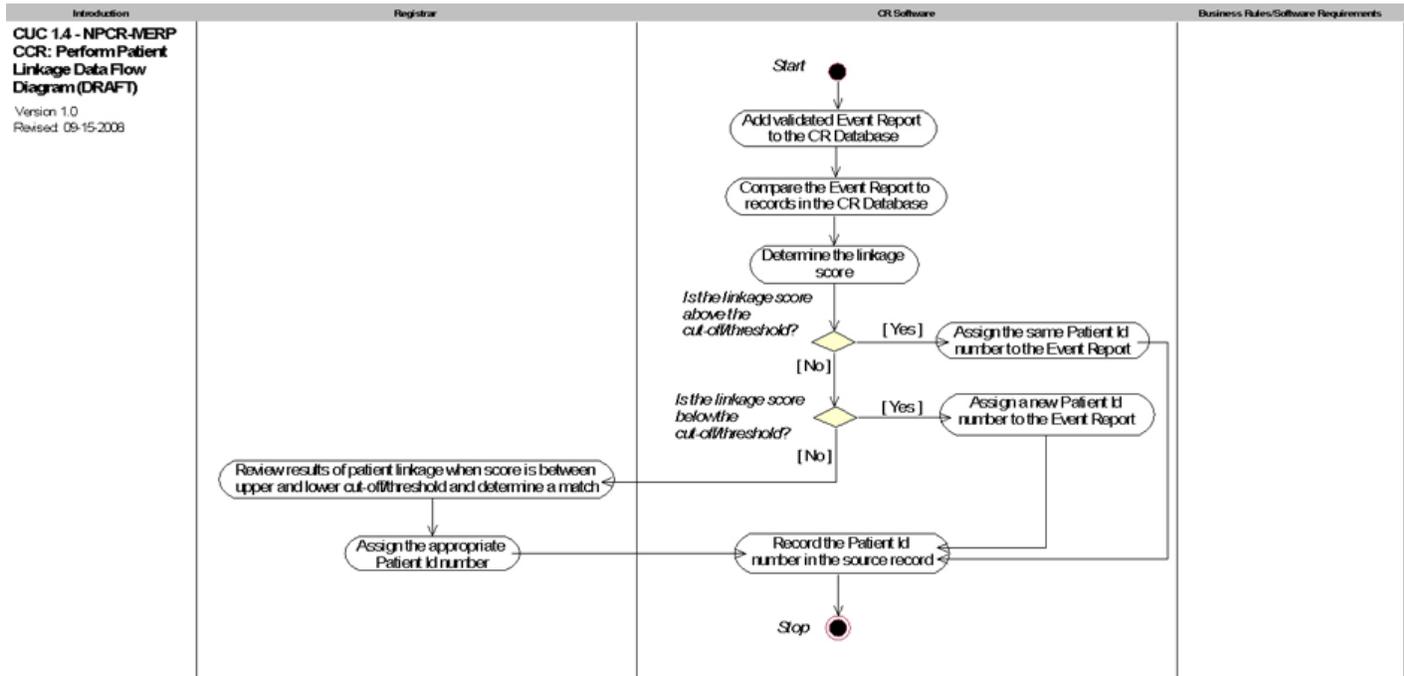
Appendix A: Perform Patient Linkage Workflow Diagram

NPCR-AERRO CCR: Perform Patient Linkage Workflow Diagram (DRAFT)

Version 1.0
Revised 9-12-2008



Appendix B: Perform Patient Linkage Data Flow Diagram



Appendix C: References for Probabilistic Record Linkage

Allen M, McDavid K, Laurent A, Yin D, O'Connor L. "Data Record Linkages to Improve Follow-up." Presentation at CDC Cancer Conference, Atlanta, Georgia, September 2003.

Belin TR, Rubin DB. [A method for calibrating false-match rates in record linkage.](#) *Journal of the American Statistical Association* 1995;90(430):694–707.

Dempster AP, Laird NM, Rubin DB. [Maximum likelihood from incomplete data via the EM algorithm.](#) *Journal of the Royal Statistical Society Series B (Methodological)* 1977;39(1):1–38.

Fellegi IP, Sunter AB. [A theory for record linkage.](#) *Journal of the American Statistical Association* 1969;64(328):1183–1210.

Jaro MA. [Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida.](#) *Journal of the American Statistical Association* 1989;84(406):414–420.

North American Association of Central Cancer Registries Best Practices Workgroup (eds). "Resolving Death Clearance Issues, 2002." *Procedure Guidelines for Cancer Registries, Series V.* January 2003.

Winkler WE. [String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage.](#) *Proceedings of the section on Survey Research Methods, American Statistical Association* 1990;354–359.

Winkler WE. [Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage.](#) *Proceedings of the Section on Survey Research Methods, American Statistical Association* 1988;667–671.

Appendix D: Matching Methods Used in Link Plus

Link Plus uses the following nine matching methods. In addition to the exact matching method, several approximate matching methods find partial, approximate, or fuzzy matches, and generate match values other than “yes” or “no.” These matching methods use partial matching, value-specific matching, or both, and are customized for the content of specific data items or types.

Tip: We recommend using the matching method defined for each type of variable. For example, the date matching method is recommended for date variables, and the last name matching method is recommended for the Last Name variable.

Exact. A case insensitive character-for-character string comparison method. The result is either yes or no.

Last name and first name. These matching methods use partial matching, value-specific matching, and NYSIIS phonetic code to account for minor misspellings and hyphenated names.

- For a hyphenated name, these name matching methods compare each substring separated by the hyphen with the other name of the comparison pair.
- If a comparison pair has the same name, the name frequency is included when computing the weight of this pair. A common name has a low weight, and a rare name has a high weight. By default, name frequencies are derived from File1; however, Link Plus allows you to use name frequencies from 2000 United States Census data or 2000 National Death Index data.
- If a comparison pair has different first names, the first name matching method uses a nickname file included in the program to match names if one of the names is considered a nickname.
- If the names don’t match after nickname checking, partial matching is tried. Partial matching is based on the *Jaro-Winkler metric* (see below), which is used widely to measure the similarity between two names. A name match status of yes or no is determined according to whether the similarity score is greater than a threshold value, or less than and equal to the threshold value. For a hyphenated name, the name matching methods compare each substring separated by the hyphen with the other name in the comparison pair.

The Jaro-Winkler metric measures the agreement between two strings by—

- Computing the string length.
- Finding the number of common characters in the two strings.
- Finding the number of transpositions between the two strings.

Common characters are defined as any same characters within half the length of the shorter string. *Transposition* is defined as a character from one string that is out of order with the same character from the other string.

Winkler enhanced the Jaro string comparator by assigning increased value to agreement on beginning characters of a string, since the fewest errors typically occur at the beginning of a string. The formula for the basic Jaro string comparator is—

$$\Phi = (c/l_1)W_1 + (c/l_2)W_2 + ((c - n_t)/c)W_t$$

Where

W_1 = weight associated with string in the first file

W_2 = weight associated with string in the second file

W_t = weight associated with transpositions

l_1 = length of string in first file

l_2 = length of string in second file

n_t = number of transpositions

c = number of common characters

The number of transpositions is calculated as follows—

- The first common character on one string is compared to the first common character on the other string. If the characters are different, half of a transposition has occurred.
- The second common character on one string is compared to the second common character on the other string, and on until the end of one name is reached.
- The number of mismatched characters is divided by two to yield the number of transpositions.

Middle name. This matching method accounts for occurrence of the middle initial only versus the full middle name.

Social Security number. This matching method incorporates partial matching to account for typographical errors and transposition of digits. The SSN matching method also enables a match between a nine-digit Social Security number in one file and a four-digit Social Security number in the other file. If the last four digits of the nine-digit number are the same as the four-digit number, the comparison pair receives a higher score.

Date. This matching method incorporates partial matching to account for missing month or day values. It compares the day, month, and year components of two dates.

- If all three components are the same, the comparison pair gets a high weight.
- If the year and month are the same but the day is missing, the weight is lower.
- If the year is the same but the month and day are missing, the weight is still lower.
- If the values are not missing, the day and month are checked for transposition.

Value-specific (frequency-based). Intended for advanced users, this matching method sets weights for matching values based on their frequencies in the files being compared. A match on a common value gets a low weight, while a match on a rare value gets a high weight. For example, if this matching method is applied to the Race variable in a file in which most records have a value of 01 (white), a match on value 01 would get a lower weight than a match on value 03 (American Indian).

Generic string. This matching method uses partial matching to account for typographical errors. It uses an edit distance function (Levenshtein distance) to compute the similarity of two long strings. The *edit distance* is defined as the minimum number of operations (insertion, deletion, or substitution of a single character) needed to transform one string into the other.

ZIP Code. The ZIP Code matching method enables the match between a nine-digit ZIP Code and a five-digit ZIP Code. If the first five digits of the nine-digit ZIP Code are the same as the five-digit ZIP Code, the comparison pair gets a high weight.

Use Case Administrative Information

1. Use Case History

None.

2. Created By

- NPCR-AERRO Central Cancer Registry Workgroup
- NPCR-AERRO Technical Development Team

3. Date Created

November 6, 2007

4. Last Updated By

SJ

5. Date Last Updated

September 22, 2011

Revision History

Name	Date	Reason for Changes	Version
CCR Workgroup	11/13/07	Reviewed and revised steps	0.02
WKS, MA	7/15/08	Updated use case	0.03
MA	7/16/08	Added actors and description	0.04
WKS	9/4/08	Added appendices C and D	0.05
WKS	9/10/08	Added Purpose content	0.06
WKS, SJ	9/22/11	Final review	1.0