Record Linkage for Registries: Current Approaches and Innovative Applications

Rich Pinder
Los Angeles Cancer Surveillance Program
rpinder@usc.edu

Nelson Chong
Surveillance Unit
Cancer Care Ontario
Nelson.Chong@cancercare.on.ca

Presented at the NAACCR Informatics Workshop
June 9, 2002

Introduction

- One of the primary functions of a cancer registry is to bring together information describing the same individual from a variety of data sources.
- Humans can conduct record linkage manually by visually comparing records from two separate sources.
- Approach becomes time consuming and tedious for a cancer registry
  - large volume of records cause manual methods to become inefficient and unworkable

Introduction **(cont'd)**

- **Multiple notifications of the same cancer likely from use of multiple sources of information**
  - **efficient record linkage procedures on same individual very important**
  - **failure in record linkage process results in missed cases and /or duplicate registrations**
- **Technological advances in computer systems and programming techniques**
  - **economically feasible to perform computerized record linkage between large files quickly and reasonably accurately**

Why Link Records?

1) Registry Operations

Because you have a Master List and wish to add new names to it.

List of Names

Add to list?

Hardy

Hardie
Harding
Mitchell
Ogilvie
Simpson

Already in list?

2) Research Linkages

Because you have two lists and wish to compare them.

List of workers

Baker
Dow
Fry
Willis
York

Which workers developed cancer?

List of cancer patients

Cook
Francis
Martin
Sanders
Willis

Objectives

- Introduce record linkage software packages
- Discuss methods to incorporate linkage into production database systems
- Provide examples of record linkage projects using central cancer registry data (Help !!!)
- Provide theoretical and practical overview of record linkage concepts

Objectives

- **Introduce record linkage software packages**
- **Discuss methods to incorporate linkage into production database systems**
- **Provide examples of record linkage projects using central cancer registry data (Help !!!)**
- **Provide theoretical and practical overview of record linkage concepts**

Record Linkage Software:

SSA-Name3
Linkpro
GRLS
DataFlux
SuperMatch

Toronto, June 2002

Record Linkage Requirements

- ✓ **Record linkage methodology**
  - ✓ **Fellegi-Sunter model?**
- ✓ **Generalizability functionalities:**
  - **2-file linkage / 1-file (internal)**
  - **batch / interactive processing**
- ✓ **Fit to hardware and software environment**
- ✓ **Readiness for use**
- ✓ **Cost and simplicity**
- ✓ **Maintenance and support**

Other software features to consider:

- **Standardization of names and addresses**
- **Interactive review of "possible" candidate matches**
- **Report generation**
- **Extraction of linked data**

a) SSA-Name3 / DCE

- **Search Software America (SSA)**
  - **supplies name search and matching software tools to end user developers, software houses and systems integrators.**
  - **Web site:** www.searchsoftware.com
- **Sets of software components for building a custom record linkage system – Interfaces into your existing systems**
- **Data Clustering Engine – standalone product**
- **Platform availability:**
  - **MVS, Unix, Windows**

a) SSA-Name3 *(cont'd)*

- **Used by companies with large databases to generate keys for efficient database searches**
- **Developers can create either a batch system or an interactive system from the software modules.**
- **Linkage can be performed directly against the database (interactive process)**

a) SSA-Name3 *(cont'd)*

- **SSA-Name3 does not use the Fellegi-Sunter record linkage methodology**
  - **treats the record linkage problem like a database search problem**
  - **software uses compressed, fixed-length 5-byte keys, based on an enhanced name coding system (NYSIIS), to efficiently locate potential matches**
  - **creates the most productive key for search purposes**
  - **using these keys, a system can be built to establish candidate sets for linkage.**
- **Costs start at $40,000, more for DCE.**

b) Linkpro

- **Infosoft Incorporated - Manitoba, Canada**
  - << web site inactive – still available ??? >>
  - **E-mail:** andre.wajda@mctrf.mb.ca
- **Written for researcher/analyst who uses SAS**
  - **system links records where no unique identifiers exist**
  - **consistent with SAS procedures**
- **Integrated SAS application system (macro) for both deterministic and probabilistic record linkage (applies Fellegi-Sunter model)**

b) Linkpro *(cont'd)*

- Calculates and applies probabilistic weights to estimate the likelihood that a pair of records from separate files corresponds to the same person.
- Converts names to SOUNDEX code to overcome spelling and pronunciation problems
- Runs on mainframe, mini, workstation or PC that has SAS version 6.07 installed
- Cost: $1,900 CDN ~ $1,300 US

c) GRLS

- **Generalized Record Linkage System (Statistics Canada)**
  - E-mail:

    Evelyn Perkins Evelyn.Perkins@statcan.ca (Technical)

    Martha Fair Martha.Fair@statcan.ca (Research)
- **Based on Fellegi-Sunter probability model**
- **Linkage operation broken into three steps:**
  - Search - **comparison rules and associated linkage weights, database of potential matches created**
  - Decide - **potential matches divided into sets of possible and definite matches**
  - Group - **records belonging to the same person are grouped together**

c) GRLS *(cont'd)*

- Allows batch (background) or interactive linkage
- Allows concurrent users for each linkage project
- Handles both one-file (internal) and two-file linkages
- Site license – price includes 5 day training
- Current version 4.0
- Platform Specifications:
  - » uses client-server architecture (Unix server, or mainframe)
  - » requires ORACLE RDBMS version 8.04, with SQL*PLUS, PL*SQL, PRO/C, FORMS & Graphics 6i runtime
- Costs: $30,000 CDN ~ $20,000 $US

d) DataFlux

- **DfPowerStudio 4.3 standalone application**
- **BlueFusion SDK – developer toolkit**
- **http://www.dataflux.com**
  - E-mail: Scott Barrett <scott.barrett@dataflux.com>
- **Deterministic model**
- **A SAS Company** **"DataFlux is a wholly owned subsidiary of SAS Institute"**

d) DataFlux (*cont'd*)

- **Interfaces with ODBC compliant database**
  - **Platform availability:**
    - » **Win NT/2000 DfPowerStudio;**
    - » **Win NT/2000 & Unix BlueFusion**

      Requires SAS
  - **Costs: $15,000- $24,000 US (depending on modules needed)**

e) SuperMatch

- Current version of Matt Jaro's AutoMatch
- Recently (5/02) acquired by Ascential Software
- http://www.ascentialsoftware.com/
- Generalized record linkage program based on Fellegi-Sunter model
  - » features both internal and two-file linkage capabilities
  - » designed for linking in multiple passes where the unlinked records from each pass proceed to the next pass (due to errors in the blocking variables)
  - » Excellent set of comparator operations
- SuperStan - available companion program used for standardization of name and address information

e) SuperMatch *(cont'd)*

- Files to be linked can be in ASCII text files, or in DBASE format
- Primarily used in batch mode
- RealTime (interface libraries) available too
- GUI client interface wraps the old familiar command line tools. Nice for new users – overkill for experienced users
- Upcoming Version (4.1) to have clerical review module reinstated

- Platform availability:
  - » Win NT/2000, Unix, Mainframe
- Costs: in negotiation !

Objectives

- Introduce record linkage software packages
- **Discuss methods to incorporate linkage into production database systems**
- Provide examples of record linkage projects using central cancer registry data (Help !!!)
- Provide theoretical and practical overview of record linkage concepts

- **Real Time environment – desirable**
  - Mimics work flow
  - Time/sequence advantage over Batch
- **Integrated vs symbiotic**
  - Sophistication vs ease of implementation
  - Can your Database environment sustain?

- **'Home Grown' Fine for production ?**
  - Simplified algorithm (Deterministic ok?)
  - Requires increased Database index/keys resources?
  - 80/20 rule – will it suffice ?
- **Third Party products advantages**
  - Better algorithms ? (Probabilistic, Complex comparators)
  - Easier to document and defend?
  - No maintenance
  - Concurrency issues

Objectives

- Introduce record linkage software packages
- Discuss methods to incorporate linkage into production database systems
- **Provide examples of record linkage projects using central cancer registry data (Help !!!)**
- Provide theoretical and practical overview of record linkage concepts

- **Production**
  - **Follow up:**
    - Mortality: State vital stats; SSA DMF;
    - Voter Registration;
  - **Work Process Flow:**
    - Pathology review
    - New case additions
    - Unduplication
- **Research**
  - **Incidence:**
    - Cohort studies
    - Aids linkages
    - Worker effects – Aircraft workers

Objectives

- Introduce record linkage software packages
- Discuss methods to incorporate linkage into production database systems
- Provide examples of record linkage projects using central cancer registry data (Help !!!)
- **Provide theoretical and practical overview of record linkage concepts**

Topics to consider

- Code standardization
- File Standardization & File review
  - Look for problems/undocumented issues in data
  - Is coding consistent
  - Review data manually – beware of formatting errors
  - How much missing data ?
  - Know accuracy of elements

Deterministic and Probabilistic Record Linkage Methods

Toronto, June 2002

"Exact Match" / *Deterministic* Linkage

- Simpler method of matching.
- Records agreeing "exactly" within an individual data field or a group of common fields between records.
- Approach relies on files having unique identifying information
  - health insurance number, social security number, surnames, given names
    - »minimal amount of missing or erroneous information

"Exact Match" / *Deterministic* Linkage

- Primary advantages:
  - **technique brings together record pairs very efficiently, simply by sorting both files using a common unique identifier as the key field.**
  - **can be successfully applied when accurately recorded unique personal identifying information is available**

"Exact Match" / *Deterministic* Linkage

- Primary disadvantages:
  - absence / incompleteness / inaccuracy of key identifying variables
    - » e.g., inconsistencies from record to record in the accuracy of surnames, given names and other identifiers, such as birth date.
  - spelling and transcription errors at time of data collection
  - use of nicknames and proper names used interchangeably; name changes over time (marriage/adoption)

"Exact Match" / *Deterministic* Linkage

- Develop rules based on variables present on both files e.g., matches if any of these conditions are met:
  1. same surname, 1st name, ID#, date of birth or
  2. same surname, 1st name, date of birth or
  3. same surname, 1st name initial, ID#, age, etc.
  - Note: there are 2^n possible patterns of agreement and disagreement on *n* fields:
    - » e.g., 10 fields = 2^{10} = 1,024 possible combinations of fields agreeing and disagreeing!

"Exact Match" / *Deterministic* Linkage

- This doesn't account for missing values and partial agreements.

- Specialized code for deterministic combinations often takes years to develop and never quite fulfills its goals. In addition, flexibility is lost.

Probabilistic Record Linkage

- **Recommended over traditional deterministic methods (i.e. exact matching) methods when:**
  - *coding errors*, *reporting variations*, *missing data* **or** *duplicate records* **encountered by registry**
- **Estimate probability / likelihood that two records are from the same person versus not**
- **Frequency Analysis of data values involved (and IMPORTANT)**

Probabilistic Linkage ***(cont'd)***

- **Landmark papers in computerized probabilistic record linkage by several Canadians in 1960s and 1970s (Fellegi & Sunter, Newcombe, Howe)**
- **Statistics Canada (in collaboration with NCIC) - developed the Generalized Iterative Record Linkage System - GIRLS (based on Fellegi-Sunter model)**
  - **Details in: Newcombe HB. Handbook of Record Linkage. Oxford University Press, 1988**

Probabilistic Linkage ***(cont'd)***

- **Frequency Analysis – examples:**
  - **How common is the surname 'Takaharu' in the Northern Texas Regional Cancer Registry?**
  - **How common is the surname 'Takaharu' in the Tokyo Cancer Registry ?**
  - **If you've got an 'iffy' match – and the Surname is 'Rumplepinder' – you likely to take it ?? (say ssn is missing, and mo/day of birth is wrong)**
  - **If you've got the same 'iffy' match – and the Surname is 'Jones' ???**

Probabilistic Linkage ***(cont'd)***

- **Frequency Analysis – examples:**
  - **You're matching your Cancer file with the Mortality file. What are the impacts of a pair of 'John M Smith' matching with month/yr agreement on birth of 10/23….. Vs the same scenario but an agreement of birth of 10/79**

- **This is a HUGE component of probability**

Probabilistic ***Linkage (cont'd)***

- **Formalization of intuitive concepts regarding outcomes of comparison of personal identifiers**
  - agreement **argues** for linkage **and** ***disagreement against*** **linkage**
  - **partial agreement is less strong than full agreement in supporting linkage**
    - **some types of partial agreements are stronger than others (e.g., truncated rare surname vs residence county code)**

Probabilistic ***Record Linkage (cont'd)***

- **Agreement on an uncommon value argues more *strongly* for linkage than a common value (e.g., surname Drazinsky vs Smith)**
- **Agreement on a more specific attribute argues more strongly for linkage than agreement on a less specific one (e.g, SSN # vs sex variable)**
- **Agreement on more attributes, disagreement on few, supports linkage**

Probabilistic **Record Linkage (cont'd)**

- Blocking:
  - **probabilistic linkage step that reduces the number of record comparisons between files**
  - **records for the two files / single file to be linked partitioned into mutually exclusive and exhaustive blocks**
  - **comparisons subsequently made *within* blocks**
  - **implemented by "sorting" the two files by one or more identifying variables**

Probabilistic Record Linkage ***(cont'd)***

- **Once comparisons within blocks are made:**
  - *weight* **calculated for each field comparison, and total weight derived by summing these separate field comparisons across all fields that have identifying value**
    - **»e.g., surname, given names, birth date**
- **Define thresholds for automatically accepting and rejecting a link**
  - **gray area / marginal links reviewed manually**

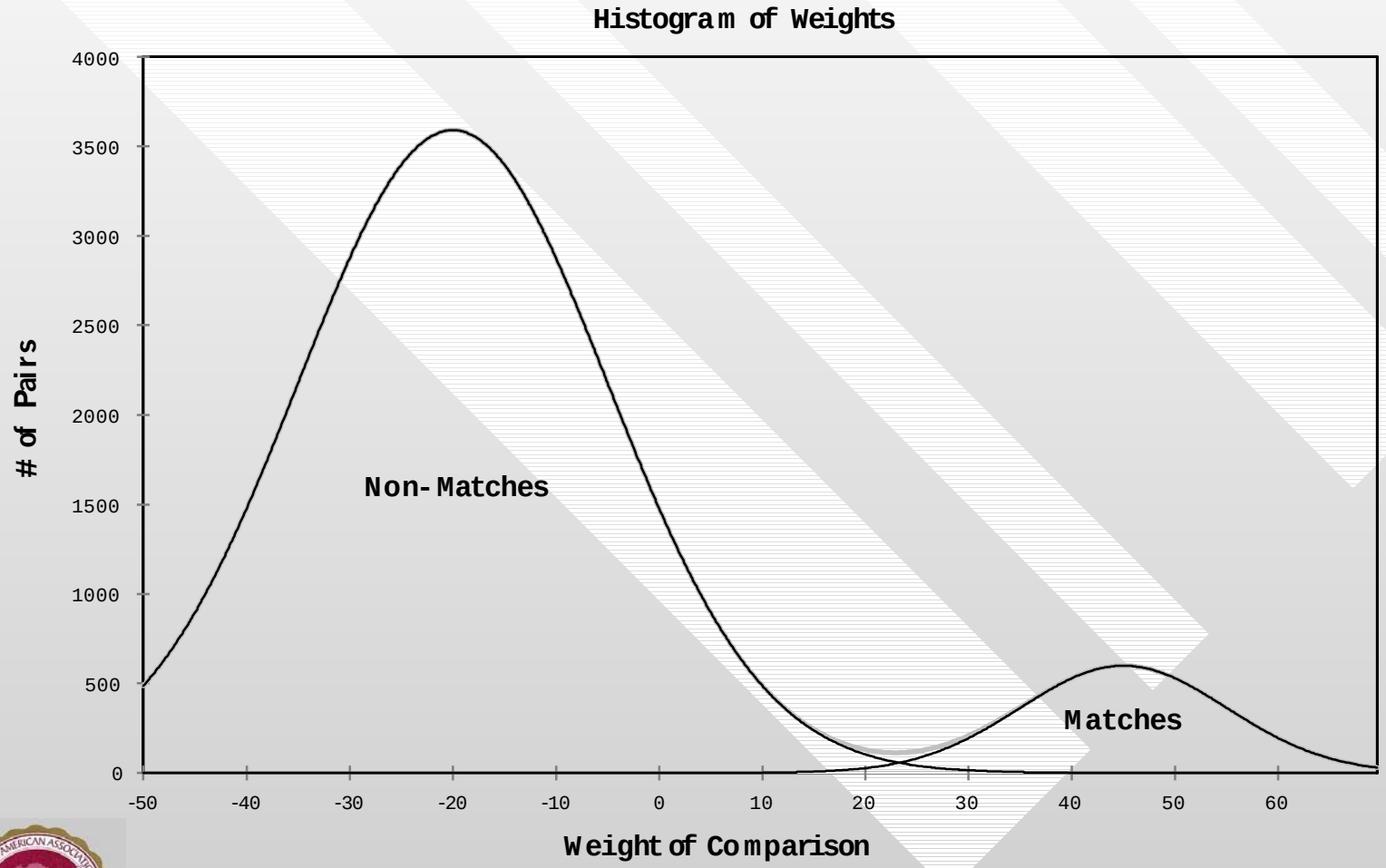
Definition of Weight (Fellegi-Sunter model)

- Each variable / field has an agreement and a disagreement weight associated with it.
- The agreement weight is log (m/u).
- The disagreement weight is log ((1-m)/(1-u))
- m is the probability that a field agrees given a correctly matched pair (measures the reliability of a field).
- u is the probability that a field agrees given a non-matched pair (ie, chance of accidental agreement)
- Logarithms are to the base two.
- The agreement weight is applied to the field if it matches in the record pair being examined, else the disagreement weight is applied.

Discrimination

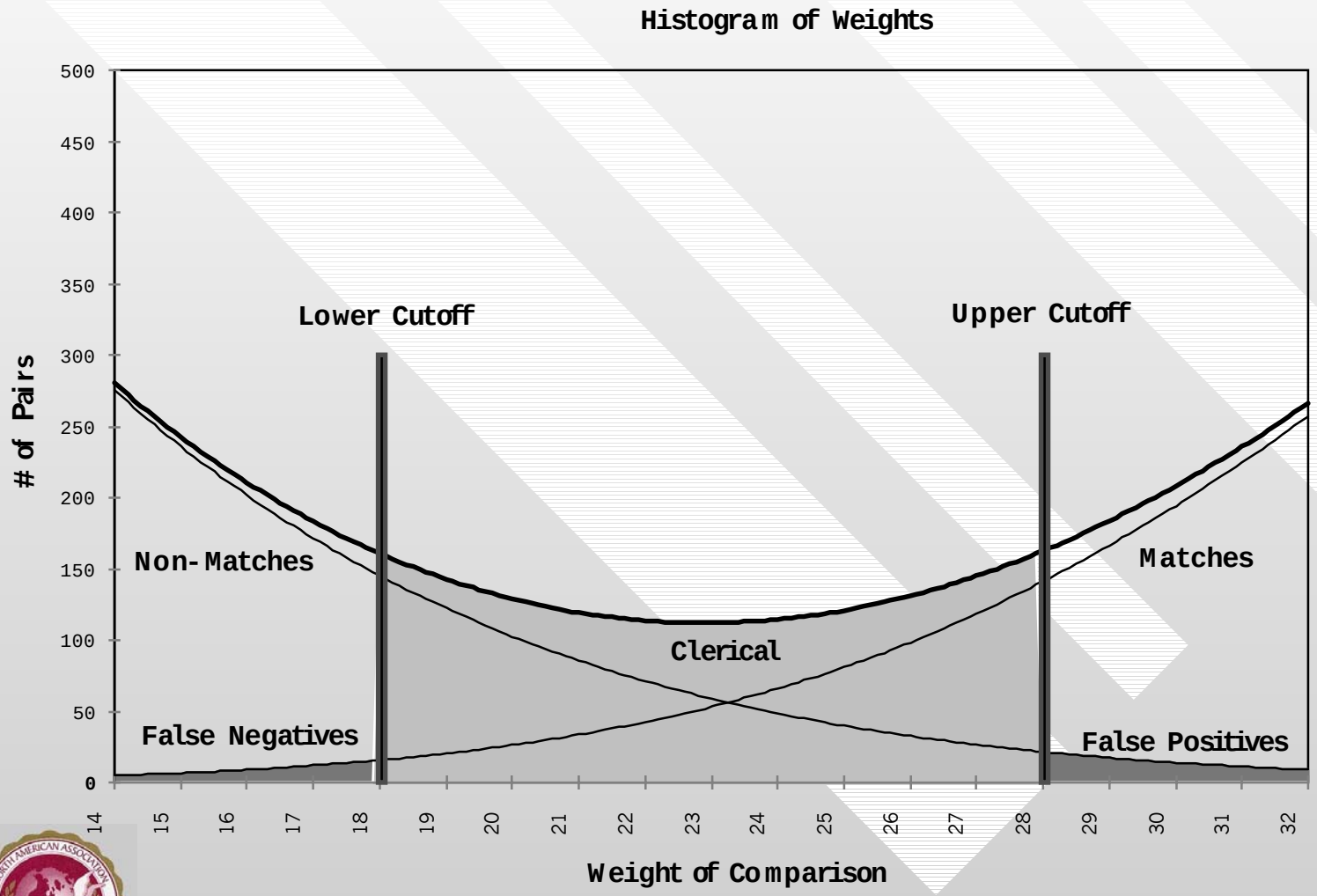
- It is the difference in the distribution of the weights for unmatched and matched pairs that enables one to discriminate between matches and non-matches.
- The more fields are available for matching, the bigger this difference will be and more reliable matches will result.

Distribution of Weights



Source: MatchWare Technologies, Inc., Burtonsville, MD, USA (1995)

Detail of Histogram



Source: MatchWare Technologies, Inc., Burtonsville, MD, USA (1995)

Probabilistic Linkage ***(cont'd)***

- **Usually we try to triage possible links on the basis of calculated likelihood. Weights (*w*) are derived to determine:**
  - **likely to belong to same individual (*w* > Upper Threshold, W_U)**
  - **uncertain if belong to same individual - "gray area" (W_L < *w* < W_U)**
  - **unlikely to belong to same individual (*w* < Lower Threshold, W_L)**

Conclusions and Recommendations

- **Commercially-available computerized record linkage programs can help reduce the cost and increase the scope and scale of case finding for cancer registries**
  - **long-term follow-up**
  - **cohort research studies**
- **Record linkage skills should be more widely distributed**
  - **all central cancer registries maintaining databases need these skills to perform internal linkages and to identify new cases**