Data Linkage for Cancer Control
Record Linkage Methodologies

Rich Pinder

Los Angeles County

Cancer Surveillance Program

RPinder@usc.edu

http://www-hsc.usc.edu/~rpinder

(Google for Rich Pinder)

Presented at the CDC's 2003 Cancer Conference

September 15th 2003

Data Linkage [aka Record Linkage]

So…. How Do We Do It ??

- **Program it Yourself – ‘In House’**
- **Purchase (acquire) a software solution**

- ‘Home Grown’ Systems
  - Simplified algorithm
  - Requires increased Database resources
  - Black Box story

- Third Party products
  - Better algorithms
  - Easier to document and defend
  - No maintenance

Topics to consider for ANY Linkage

- **File Standardization & Preliminary File review**
  - **Look for problems & ‘surprises’ in data**
  - **Is coding consistent ?**
  - **How much missing data ?**
  - **Know accuracy of elements**
  - **Review data VISUALLY – beware of formatting errors**

Deterministic and Probabilistic Record Linkage Methods

Atlanta   September, 2003

IF all datasets we wanted to link had variables that:

- Included ALL demographic info
- 100% Accurate
- 100% Complete

..linkage wouldn't be difficult

(and this discussion of different methodologies would NOT be important) n **Deterministic**

- Records match exactly for specified data items or variables.
- Pre determined ‘Rules’ define which variables are compared

n **Probabilistic**

- Estimates probability that records match using mathematical formulas
- Weights are calculated to select BEST matches

Deterministic Record Linkage

Atlanta   September, 2003

Deterministic Record Linkage

- **Simpler method of matching**
- **Records agreeing “exactly” within an individual set of fields or variables**
- **Works best with high quality data**

Deterministic Record Linkage

- **technique brings together record pairs very efficiently, simply by sorting both files using common identifier(s), which is the notion of a 'Key'.**

(Keys are associated with the concept of Indexing or Sorting)

Deterministic Record Linkage

- Sample 'Keys' (matches):
  1. SSN + surname + given name + date of birth
  2. SSN + surname + given name
  3. surname + given name + date of birth
  4. SSN

(Any make you nervous?)

Deterministic Record Linkage

- **Doesn't account for missing values and partial agreements.**
- **Perfecting complex 'Keys' often takes years**
- **To get acceptable results, must do LOTS of clerical review (The Human Touch!)**

Deterministic Record Linkage

Pseudo logic – to build it at home:

- **Sort both files on ‘Key’**
- **Start at record 1 in both files**
- **Step through each file looking for Keys that match (0, 1, or many)**

Probabilistic Record Linkage

Atlanta   September, 2003

Probabilistic Record Linkage

- **Probability definition (from dictionary.com):**

- **1: a measure of how likely it is that some event will occur; "what is the probability of rain?";**

Probabilistic Record Linkage

- **Recommended over simpler, deterministic methods, especially when:**
  - *coding errors*, *reporting variations*, *missing data* **or** *duplicate records* **encountered by registry**
- **Estimate probability (likelihood) that two records are the same person**
- **Frequency Analysis of data values in both files is IMPORTANT**

Probabilistic Record Linkage

- **Frequency Analysis**
  - **The counts of individual values of the variables**

- **Frequency Analysis – situations:**
  - **Rumplepinder vs Smith**
  - **How common is the surname ‘Takaharu’ in the Northern Texas Regional Cancer Registry?**
  - **How common is the surname ‘Takaharu’ in the Tokyo Cancer Registry ?**

Probabilistic Record Linkage

- **Agreement on an uncommon value argues more *strongly*** for **linkage than a common value**
- **This is a HUGE component of probabilistic record linkage**

Probabilistic Record Linkage

- Blocking (aka Passes):
  - **Efficiency step that reduces the number of record comparisons between files**
  - **Breaks project into manageable parts**
  - **GREAT analogy: Blocking is like separating your socks into piles based on Color, BEFORE you sort them.**
  - **Typically 3 or more blocks in a project**

Probabilistic Record Linkage

- Complex Comparators
  - **Can detect Sub-strings, random inserts/deletes, transpositions in character data**
  - **Numbers matched with tolerences (+-)**
  - **Prorated weights are assigned**

Probabilistic Record Linkage

The Matching process can be summarized as follows:

- **The project broken is down into blocks or passes – to make it more efficient**
- **Within a given block, all match variables are compared and weights are computed using mathematical probability based assignment.**
- **Cutoff values are applied to the weights – above a certain level, EVERYTHING is a match. Below a certain level, NOTHING is a match. In between are records needing CLERICAL REVIEW**

In Summary

- **Deterministic approach & in house development sometimes easier and cheaper, but yields less success than Probabilistic**
- **Probabilistic approach uses blocking to improve efficiency**
- **Probabilistic systems often use complex comparator operations to increase match rate**
- **Clerical Review important component of any system**
- **Probabilistic affords ‘smart’ clerical review:**

  - larger number of true matches

  - smaller number of clerical reviews

Two Additional Record Linkage Resources for Cancer Registries

Atlanta   September, 2003

Social Security Administration

Service to Epidemiological Researchers to Provide Vital Status Data on Subjects of Health Research

Atlanta   September, 2003

- http://www.ssa.gov/policy/about/epidemiology.html

  (or search the SSA site on ‘epidemiology’)
- **Straight forward application process – tailor made for Cancer Registries**
- **Linkage performed as service by SSA**
- **Deterministic (simple) linkage – SSN required**
- **2003 Costs:** $ 0.17 for 1-20,000 records

  $ 0.013 20,000+ records

Data Contains:

- Numident (includes SSA's death master file)
- Master Beneficiary Records (MBR) for Title II beneficiaries (i.e. Medicare)
- Social Security Record (SSR) for Title XVI beneficiaries (i.e. Welfare / SSI)
- Master Earnings File (MEF) (i.e. FICA payments)

National Death Index

National repository of all 50 States' Vital Statistic Data

Atlanta September, 2003

- Expensive (especially for a true ‘search)
- Great for augmenting Cause of Death for known decedents
- No ability to review actual data for matches (forced to accept deterministic rules ‘blind’)

- ndi@cdc.gov
- http://www.cdc.gov/nchs/r&d/ndi/ndi.htm

  (go to CDC and search for NDI)

Resources:

- Tools to format data (Python)
- Readings on Record Linkage
- Detailed contact info for Linkage Data Resources

Search Google for 'Rich Pinder'

RPinder@usc.edu

Atlanta   September, 2003